Secondary structure in trinucleotide repeat DNA *in vivo*

John M. Darlow

Thesis presented for the degree of Doctor of Philosophy
University of Edinburgh
1999

Declaration

I hereby declare that the work described is my own, and that this thesis was composed by myself, unless otherwise stated.

John M. Darlow

March 1999

Abstract

Since 1991 an increasing number of human inherited disorders has been found to be due to expansion of naturally occurring polymorphic trinucleotide repeat tracts within the respective genes. At the time this project was begun, in 1994, all of the disorders of this type so far discovered were due to expansion of d(CAG)·(CTG) repeats or d(CGG)·d(CCG) repeats and it was suggested by many that these sequences might be more prone to mutation than other repeated sequences because their single strands might form unusual DNA secondary structures, particularly imperfect hairpins, through self-complementarity.

Various studies have established that these structures do form *in vitro*. In the meantime one trinucleotide repeat expansion disorder has emerged with a different sequence, $d(GAA)_n \cdot d(TTC)_n$. This sequence can form triplexes but, also, $d(GAA)_n$ single strands are known to form secondary structures, thought to be tetraplexes, and dGNA has been shown to be capable of forming a very tight hairpin loop. This project has investigated secondary structure formation by these and other sequences *in vivo*. Earlier work in this laboratory showed that alterations in the central sequence of a palindrome in bacteriophage λ affect the ability of the palindrome to inhibit plaque formation. Sequences known to form tight hairpins lead to the formation of smaller plaques.

By inserting different numbers of different trinucleotides into the centre of a long palindrome it has been possible to investigate their tendencies to form hairpins *in vivo* in any particular alignment and with odd or even numbers of repeat units in the hairpin. It is shown that with d(CAG)·d(CTG) repeat tracts there is a markedly greater tendency to form hairpins with even numbers of repeat units than with odd numbers whereas d(GAC)·d(GTC) repeats (which are rare, short, and have not been found to expand) show no such alternation despite having the same base composition. $d(CAG)_2 \cdot d(CTG)_2$ behaves like DNA sequences known to form two-

base loops *in vitro* suggesting that one or both of the strands may also form a compact and stable loop.

$d(CGG)_2 \cdot d(CCG)_2$ also produces very small plaques but beyond $d(CGG)_3 \cdot d(CCG)_3$ the pattern is different from that of d(CAG)·d(CTG) repeats. It seemed likely that this might be because d(CGG)·d(CCG) repeats have more than one possible alignment in which they could self-anneal. Further investigation has shown that while even-membered hairpins are preferred in the frame d(CGG)·d(CCG), hairpins with odd numbers of trinucleotides are more stable in the frame d(GGC)·d(GCC).

A disadvantage of the phage construct is that the orientation of the inserted sequence cannot be predetermined and cannot be ascertained afterwards because it is not possible to sequence across the palindrome. A new phage λ derivative has been constructed which allows not only the predetermination of insert orientation but the introduction of inserts with random central sequences for screening for ones producing tight hairpin loops. Furthermore, sequencing of the centre is possible if a single-base-pair asymmetry is used in the insert. With this construct it is shown that orientation does not affect the *in vivo* test of hairpin-forming potential. It is also shown that dGAA appears to form a tight hairpin loop *in vivo* as it does *in vitro*.

A detailed review has also been made to determine the reasons why different investigators have come to different conclusions as to the nature of secondary structures formed by $d(CGG)_n$ and $d(CCG)_n$ single strands and what the likely structures might be *in vivo*.

Acknowledgements

I thank my supervisor, David Leach, for selecting me for this project and for many helpful discussions, the Medical Research Council for funding me initially and my parents for their continuing financial and moral support after the grant came to an end, Thorsten Allers for teaching me most of the techniques used in this work, including use of equipment and computer software, John Findlay and Chris Jeffree for help with setting up the image analyzers, Lynne Powell and Lucy Kirkham for instruction in Maxam-Gilbert sequencing, Markus Winter for help with compiling my electronic bibliography database and various other help, Kristina Schmidt for discussions of the work, computer tips, and passing on of references, and other members of the group and floor for various support.

Table of Contents

Abbreviations and acronyms

Ac	acetate, as in, for example, NaAc
bp	base-pairs
BAA	bromoacetaldehyde
BSA	bovine serum albumin
CCD	Cleidocranial dysplasia
CJD	Creutzfeldt-Jakob disease
CEPH	Centre d'Etude du Polymorphisme Humain
DEPC	diethylpyrocarbonate
DM	myotonic dystrophy (*dystrophia myotonica*)
DMS	dimethyl sulphate
dNTPs	deoxynucleotide triphosphates
DRPLA	dentatorubral-pallidoluysian atrophy
DTT	dithiothreitol
EDTA	disodium salt of EthyleneDiamineTetra-Acetic acid, now known as diaminoethanetetra-acetic acid
EM	electron microscope
FMR1	gene involved in fragile-X syndrome (Familial Mental Retardation)
FMR2	gene involved in mental retardation associated with FRAXE
FRAXA/E/F	fragile site A/E/F on the X chromosome
*FRA10/11*etc	fragile sites on autosomes 10, 11, *etc.*
HD	Huntington disease (*alias* Huntington's chorea)
kb	kilobase pairs
MJD	Machado-Joseph disease (now spinocerebellar ataxia type 3)
MOI	multiplicity of infection
NMR	nuclear magnetic resonance
NTM	'Normally transmitting male' (male carrier of fragile-X syndrome)

OPMD	occulopharangeal muscular dystrophy
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PEG	polyethylene glycol
p.f.u.	plaque-forming units
PSQ	plaque size quantification
RED	repeat expansion detection method of Schalling *et al.* (1993)
r.p.m.	revolutions per minute
SBMA	spinal and bulbar muscular atrophy
SCA1/2/etc.	spinocerebellar ataxia type 1/2/etc.
SPD	synpolydactyly
TEMED	N-N-N'-N'-tetramethyl-1,2-diaminoethane
Tris	2-amino-2-(hydroxymethyl)-1,3-propandiol
Triton X-100	oxyphenoxypolyethoxyethanol
VNTR	variable number tandem repeat
v/v	volume per volume
w/v	weight per volume

Chapter 1

Introduction

The discovery of repeat expansion as a cause of inherited disorders

Arrays of tandem repeats, polymorphic in the number of repeat units (variously known as variable number tandem repeats, VNTRs, short tandem repeats, STRs, simple sequence DNA, microsatellites and minisatellites[1]) occur throughout complex genomes and have been put to use for DNA fingerprinting, producing genetic maps, and localizing disease genes by linkage (Jeffreys *et al.*, 1985). Much interest had been shown in their evolution (Tautz & Renz, 1984; Levinson & Gutman, 1987; Jeffreys *et al.*, 1988; Epplen *et al.*, 1991) but until the beginning of this decade they had not been perceived as a cause of disease, occurring as they do, mainly in non-coding DNA. In 1991 Epplen *et al.* (1991) commented that their study was "neither particularly fashionable nor lucrative". This changed that same year with the sequencing of the mutations responsible for two human inherited disorders. The genes each contained a tract of trinucleotide repeats that was polymorphic in length and stably inherited in normal individuals but of increased length and unstable in affected individuals.

Fragile-X syndrome

The first disorder was Fragile-X syndrome. It is the commonest form of inherited mental retardation and is associated with a number of characteristic dysmorphic features. It was first described by Martin & Bell (1943) and hence was

[1] Definitions of microsatellites and minisatellites vary between authors but microsatellites are usually considered to have repeating units of 1 - 4 or 1 - 6 bp and minisatellites start above them.

originally known as the Martin-Bell syndrome. It was clearly X-linked but unusual. In the original family, two unaffected brothers had passed the gene on through their unaffected daughters to the following generations. In 1969, Lubs reported a family in which some members had a 'marker X' chromosome. There was a constriction near the end of the long arms of the chromatids. All four male family members who bore the marker X were mentally retarded. Two unaffected females also showed the marker X and one of those was an obligate carrier of the mental retardation, though the marker X was not seen in another carrier female. It was not immediately recognized that this was the same syndrome because the dysmorphic features were not evident and the original and later-reported families had not been examined cytogenetically, and it became known as the marker (X) syndrome. The constriction seen by Lubs was an example of a fragile site.

Fragile sites

'Fragile sites' are loci that appear as non-staining regions, or occasionally breaks, when the cells from which the chromosomes are prepared are exposed to particular conditions of cell culture or chemical agents (Sutherland, 1991a; Sutherland & Richards, 1995). There are now over one hundred known fragile sites on the human chromosomes and they are divided into the common sites which probably occur in all individuals, *i.e.* are a part of normal chromosome structure, and the rare sites which vary in frequency from about 1 in 20 individuals to ones seen only in a single family. They are usually inherited in a Mendelian manner and both classes are subdivided by the conditions under which they are expressed. It has been suggested that they may predispose to chromosome breakage *in vivo* and lead to an increased rate of sister chromatid exchange and recombination. The common sites at least are highly conserved in evolution. More than 50% of breaks that have occurred during chromosome evolution in primates are reported to be at or near fragile sites (refs in Knight *et al.*, 1993). The molecular basis of these sites is thus of considerable interest.

Despite their possible instability, fragile sites are usually not associated with any abnormality. Fragile-X syndrome, as the condition later became known, was the first exception noticed. The fragile site involved, now known as '*FRAXA*', is a rare folate-sensitive site (Sutherland & Richards, 1995). After Lubs' (1969) report, no other families with X-linked mental retardation were found to have marker X chromosomes until the mid 1970s (Sutherland, 1985). This turned out to be because, at around the time of Lubs' report, most cytogenetics laboratories started using newly developed culture media which yielded higher quality chromosome preparations and it was only after it was discovered that the fragile site was expressed in the old culture medium (TC199), but not in the new ones, that the condition was discovered to be common (Sutherland, 1985).

The Sherman Paradox

Analysis of pedigrees by Sherman *et al.* (1984; 1985a) detailed the very unusual nature of this X-linked dominant condition. 30% of carrier females have some degree of mental impairment and 20% of males whose X-chromosomes have the fragile site are phenotypically normal. The mentally retarded males usually do not reproduce but these 'normally transmitting males' (NTMs) pass their mutant allele on to their daughters who are also unaffected, but their grandsons are often affected. Brothers of NTMs have only about a 9% risk of being affected whereas for grandsons it is about 40% and for great grandsons about 50%. These findings became known as The Sherman Paradox yet, at the same time, the germ of the explanation had been proposed (Pembrey *et al.*, 1984). It was quite different from the explanation of Sherman *et al.* (1984) who thought that all mutations must occur in sperm and that over half of 'random' carrier females must be fresh mutants.

Deduction of the mechanism

Pembrey *et al.* (1984) pointed out that the heterozygous daughters of NTMs were never mentally retarded and either have no fragile sites or have them in very few

cells whereas by contrast in the next generation about a third of carrier females are affected with an average of 29% of cells showing the fragile site. They proposed (Pembrey *et al.*, 1984; Pembrey *et al.*, 1985; Winter *et al.*, 1985; Pembrey & Winter, 1985) that there was first a 'premutation' that did not cause symptoms but is liable to further mutation to 'the full mutation'. This second event, they suggested, was almost certainly a recombination, since it only occurred when being passed on by a female, and probably generated a duplication or a deletion. One observation of Sherman *et al.* (1984) that it did not explain was that there was an inverse relationship between the I.Q. of carrier females and the frequency of marker (X) expression in their cells. Sherman *et al.* (1985b) themselves pointed out that their observation that the offspring of obligate carrier mothers of NTMs consistently differed from those of daughters of NTMs with respect to the expression of the fragile X syndrome (*i.e.* the mothers and the daughters of NTMs, who should both carry the premutation, should have the same risk of having an affected son, but they did not) presented a strong argument against the model.

Nussbaum *et al.* (1986) noted this objection but did not reject the model; instead, he proposed a modification. Investigations by Sutherland *et al.* (1985; 1986) of the precise sensitivities of the fragile site to concentrations of DNA precursors led them to conclude that the site was almost certainly a sequence of multiple copies of a short repeat motif. Based on this work and work of their own on *FRAXA* expression in somatic cell hybrids, Nussbaum *et al.* (1986) proposed that the premutation was an amplification of the normal *FRAXA* locus, perhaps by unequal meiotic recombination. This would then be a better substrate for further unequal crossing-over, leading to further amplification and/or deletion, producing the full mutation. More amplification would increase the sensitivity of the DNA to induction of the fragile site. They further refined the model (Ledbetter *et al.*, 1986) by suggesting that there might be continuous variation in length or copy-number of the fragile-site DNA resulting in varying degrees of cytogenetic expression and a threshold for clinical manifestation. This model was supported by the observation of Warren *et al.* (1987)

that when the size of the fragile site locus was reduced by translocation at the locus in somatic-cell hybrids, the frequency of fragile-site expression was reduced.

Laird (1987) accepted that mutation must be a two-step process but rejected the rest of the model because he noted from the pedigrees of Oberlé *et al.* (1986) with two flanking marker loci that crossing-over in a carrier mother was not a prerequisite for expression of the syndrome in her progeny. Laird (1987) proposed that the fragile-X mutation was a local block to X-chromosome reactivation that occurs in a female before oogenesis. A cycle of X-chromosome inactivation and incomplete reactivation would result in local imprinting which would inhibit transcription in the region, thereby causing the fragile-X syndrome. NTMs and some heterozygous females were unaffected because their X-chromosomes had not been imprinted in a previous generation. Variable expression in females with an imprinted X-chromosome resulted from random inactivation of X-chromosomes in somatic cells. Laird saw his model as an alternative to previous model. Actually, the truth turned out to be a combination of the two.

The molecular mechanism revealed

In 1991 it was first shown that the region contained a CpG island that was methylated in affected individuals but not in NTMs or normal males (Bell *et al.*, 1991; Heitz *et al.*, 1991; Dietrich *et al.*, 1991). Then it was found that a GC-rich region was increased in length in NTMs and some females, including daughters of NTMs, and much further increased in affected individuals, who showed length variation between individuals within the same family and between cells in the same individual, indicating that the sequence was unstable in both meiosis and mitosis (Oberlé *et al.*, 1991; Yu *et al.*, 1991). Oberlé *et al.* (1991) pointed out that the alleles in the smaller range of increased size, that were not associated with methylation of the nearby CpG island or mental retardation, corresponded to premutations and the longer alleles - up to ten or more times the length of the premutations and associated with methylation of the CpG island, expression of the fragile site, and mental

retardation - corresponded to 'full mutations'. All males with fragile-X expression in ≥ 4% of cells were found to be mentally retarded, but not all such females (Oberlé *et al.*, 1991). The length of the region increased from generation to generation but only when transmitted through females (Oberlé *et al.*, 1991; Yu *et al.*, 1991).

The gene, named *FMR-1* (for Familial Mental Retardation)(Verkerk *et al.*, 1991) was found to contain a tract of d(CGG)·d(CCG) repeats with d(AGG)·d(CCT) interruptions in normal individuals (the first reported sequence was $(CGG)_{10}AGG(CGG)_9AGG(CGG)_9$ on the coding strand)(Verkerk *et al.*, 1991) and variations in length of the repeat tract in normal individuals were found to be inherited in a stable co-dominant manner (Kremer *et al.*, 1991). Further detailed study (Fu *et al.*, 1991) found a range of 6 - 54 repeats in normal individuals and 52 - > 200 repeats in premutations. Full mutations range from ~230 copies to several thousand (Fisch *et al.*, 1995). In the study of Fu *et al.* (1991) all repeats with 46 or less repeats were found to be stable and all those with more than 52 repeats to be unstable, the risk of further expansion increasing with increased length of the tract.

Thus the paradox of increasing risk with successive generations was resolved just as had been predicted. It was also found (Pieretti *et al.*, 1991) that *FMR-1* mRNA was undetectable in most affected males, suggesting that the mental retardation might indeed be caused by silencing of the gene by the methylation which followed expansion of the repeat tract. Though many authors quote the original sequencers of the gene to have discovered the repeat tract to be in the 5'-untranslated region of the *FMR-1* gene, they actually believed that it was translated as polyarginine (Verkerk *et al.*, 1991; Pieretti *et al.*, 1991; Fu *et al.*, 1991). Doubt was thrown on this by Yu *et al.* (1992) who pointed out that after 69 bp 3' to the repeat tract there followed an ATG that corresponds to the eukaryotic consensus for initiation. Furthermore, this methionine is followed by 13 out of 20 hydrophobic amino-acids characteristic of an extracellular signal peptide, so they suggested that the repeat might be in the 5'-untranslated region of the gene. This was later confirmed by Ashley *et al.* (1993).

Questions that remained were of when expansion takes place and what controls it. Oberlé *et al.* (1991) found that some males were mosaics with unmethylated premutations and methylated full mutations. They pointed out that this was difficult to reconcile with expansion taking place in oogenesis, on a previously inactive X chromosome where the *FMR1* locus was still methylated, since this would imply that secondary somatic mutation in the son resulted both in a return to the premutation and a loss of methylation. An alternative was that expansion might take place in early embryogenesis, offspring inheriting unmethylated premutations from their mothers and methylation occurring after expansion in those cells that generated full mutations. The difficulty with this is that expansion does not occur when premutations are inherited from the father.

It was subsequently shown that affected males (with a methylated full mutation in their blood) have an unmethylated premutation in their sperm, suggesting that expansion might indeed occur post-zygotically, after day 5, when germline and soma diverge, and before day 20 because twins have the same ranges of repeats as do different tissues in the same individual (refs in Bates & Lehrach, 1994). It was also shown (Sutcliffe *et al.*, 1992) that in chorionic villi of affected foetuses (with methylated full mutations and absent expression of *FMR1*) the full mutation is present but the CpG island is hypomethylated and *FMR1* is expressed, suggesting that methylation takes place in early embryogenesis. This would exclude methylation as the imprinting mechanism resulting in expansion of premutations only when inherited from the mother.

More recently, Malter *et al.* (1997) have searched gonadal tissue of full mutation foetuses of both sexes and found no evidence of premutation in germline cells and have concluded that premutation sperm result from selection of spermatogonia that have undergone contraction of the repeat tract. They also found the full mutation to be unmethylated in foetal oocytes, reinforcing the conclusion that methylation occurs after expansion. Moutou *et al.* (1997) suggested that if expansion were post-zygotic one should expect that mosaic offspring would tend to be

produced by mothers with smaller premutations while non-mosaic full-mutation offspring would tend to come from mothers with larger premutation sizes. However, they found no correlation between maternal premutation size and percentage of mosaic offspring so agreed with Malter *et al.* (1997) that mosaicism is most likely the result of postzygotic contraction of a prezygotically expanded allele.

Recognition of the significance of fragile-X

The significance of the confirmation that fragile-X syndrome was caused by an expanding repeat sequence was immediately recognized. Sutherland *et al.* (1991b, 3rd August) suggested that such heritable unstable sequences could be present in other parts of the genome and that these might explain a number of phenomena that were not well understood in terms of Mendelian genetics, including anticipation (see below), incomplete penetrance, and variable expression. They did not have to wait long. By the time their hypothesis appeared in print, the discovery that another inherited disorder was caused by expansion of a repeat tract had already been published (La Spada *et al.*, 1991, 4th July).

Spinal and bulbar muscular atrophy

Comparison of the *FMR-1* gene sequence with DNA databases (Verkerk *et al.*, 1991; Kremer *et al.*, 1991) revealed that the closest match was with the androgen receptor gene, *AR*, which has a short d(GGC)·d(GCC) repeat in the first exon coding for polyglycine (Chang *et al.*, 1988; Tilley *et al.*, 1989) and this was indeed the next gene to be found to be subject to mutation by an expanding repeat sequence but, ironically, this was not the d(GGC)·d(GCC) repeat.

Kennedy's disease (spinal and bulbar muscular atrophy, or SBMA) is an X-linked recessive form of adult-onset motor neurone disease in which affected males may show signs of androgen resistance (gynaecomastia and reduced fertility). Since the androgen receptor gene and SBMA had been mapped to the same region of the X chromosome, La Spada *et al.* (1991) examined the sequences of *AR* as a candidate gene

in SBMA families and in every affected individual they found an expansion of a d(CAG)·d(CTG) repeat tract coding for polyglutamine. The length of this tract was different in the first two published normal sequences (Chang *et al.*, 1988; Tilley *et al.*, 1989) and in particular had been found to be highly polymorphic by Edwards *et al.* (1992) whose paper had been submitted and was known to La Spada *et al.* (1991).

Edwards *et al.* (1992) found an overall range of 11 - 31 copies with different ranges and modes in different ethnic groups. La Spada *et al.* (1991) found a range of 17 - 26 in their controls and 40 - 52 in the patients. This was a much smaller range of expansion than seen in fragile-X syndrome but this reflected the different mechanism by which the expanded repeat caused the disease - here the repeated trinucleotide was coding - rather than necessarily implying a different mechanism of expansion of the repeat tract.

Myotonic dystrophy - anticipation vindicated

Early the next year another disorder was added to the list, myotonic dystrophy (Harley *et al.*, 1992a; Buxton *et al.*, 1992; Aslanidis *et al.*, 1992; Brook *et al.*, 1992; Mahadevan *et al.*, 1992; Fu *et al.*, 1992), which was also formerly known as myotonic atrophy and *dystrophia myotonica*, and has been given the initials DM (despite the fact that DM already stood for diabetes mellitus). This is an autosomal dominant disorder with a wide range of severity and age of onset. It is one of a number of disorders which had been noted to appear to become more severe and of earlier age at onset in successive generations of a family. This process had been noted in some human inherited conditions for well over a century and is known as 'anticipation' and was formerly also known as 'antedating' and 'progressive degeneration' and had a long history of dispute as to whether it was a real or perceived phenomenon (Penrose, 1948; McInnis, 1996).

Bell (1947) produced good data illustrating antedating in myotonic dystrophy and Penrose (1948) analysed it and concluded that the variability was due to modifying genes and that the appearance of anticipation was due to ascertainment

bias, in particular because he concluded that the condition was more likely to be reported if it was diagnosed in parent and child at the same time. In its mildest form, developing in middle age, cataracts may be the only feature; development earlier in adult life brings myotonia and progressive muscle weakness in addition. There may also be cardiac conduction and smooth muscle function defects, abnormal glucose response, and in males baldness and testicular atrophy. Congenital cases have respiratory distress, extreme hypotonia, muscular atrophy, feeding difficulties, mental retardation, and a high neonatal mortality.

The repeated DNA sequence was again found to be the trinucleotide d(CAG)·d(CTG) (Brook *et al.*, 1992; Mahadevan *et al.*, 1992; Fu *et al.*, 1992), but this time the CTG was on the coding strand but in the 3'-untranslated region of a gene. The gene was found to have sequence homology with protein kinase genes and was named the myotonin protein kinase gene, *MPK* (Fu *et al.*, 1992) but it may be that the functions of several genes are affected by the repeat expansion (Harris *et al.*, 1996). Subsequently the gene with the expansion was unfortunately named *DMPK* for myotonic dystrophy protein kinase by Carango *et al.* (1993) and this name has stuck. The repeat was found to be of variable length in normal individuals (5 - 30 copies in the initial reports) and to range from about 50 repeats to >6 kb (*i.e.* >2,000 repeats) in affected individuals (Brook *et al.*, 1992; Mahadevan *et al.*, 1992; Fu *et al.*, 1992), as in fragile-X syndrome, and to be highly unstable, varying between affected siblings (Harley *et al.*, 1992a; Buxton *et al.*, 1992; Aslanidis *et al.*, 1992). The length of the repeat tract was noted to be positively correlated with the severity of the condition, and negatively with the age of onset, and to tend to increase in length from one generation to the next, thereby explaining the phenomenon of anticipation (Buxton *et al.*, 1992; Mahadevan *et al.*, 1992; Fu *et al.*, 1992; Richards & Sutherland, 1992a; b; Harley *et al.*, 1992b; Tsilfidis *et al.*, 1992; Harper *et al.*, 1992; Sutherland & Richards, 1992).

This was a triumph of observation over theory, but in fairness to the mathematical geneticists of the past, it should be pointed out that when it was not

even known of what substance genes were made, let alone its structure, and when Mendel's laws were thought to be unbreakable, such a solution could hardly have been imagined. It should also be pointed out that Penrose's general conclusion that, in inherited disorders with variable age of onset, apparent anticipation is always likely to be found, was not wrong. See recent discussions of this problem by McInnis (1996) and Fraser (1997). In brief, for a condition of variable age of onset in which there is really no relationship between parents' onset ages and their children's onset ages (*i.e.* average onset ages are the same in each generation) there will appear to be anticipation even if the disease does not affect fertility because some of the children who are going to be affected late will not yet have become affected and this will lower the average age of ones who are counted. If the disorder does reduce fertility, *e.g.* if earlier onset brings more severe disease meaning less chance of becoming a parent, this will tend to raise the average age of affected people who are parents, increasing the apparent anticipation. As the majority of human disorders have variable age of onset, sorting out which ones are likely to have real biological anticipation and which ones only appear to show anticipation is a real problem though some investigators seem not to have realized this. For myotonic dystrophy, Höweler *et al.* (1989) carefully addressed all the causes of apparent anticipation and concluded that there might be real anticipation in this disease and within three years they were proved right.

Mildly affected individuals have 50 - 80 repeats and the worse affected up to >2000, but repeat length does not correlate with severity absolutely and family predisposition suggests that there *are* genetic modifiers. Normal alleles are stable. Alleles of 50 - 80 units are fairly stable but instability increases with repeat number. Repeat number sometimes decreases, usually when passed by the father, and as the repeat number in the father increases the chances that he will transmit a decrease increase (Bates & Lehrach, 1994). In one case, reduction was shown to be due to a discontinuous gene conversion event (O'Hoy *et al.*, 1993). There appears to be an upper limit of about 1000 repeats in sperm (Jansen *et al.*, 1994). It had been

observed long before the nature of the mutation was known that in the congenital cases the mutant allele was always inherited from the mother (Harper, 1989). This was shown to be because the largest expansions in DM occur on maternal transmission. When a father transmits repeats in the congenital range they are always at the lower end (Bates & Lehrach, 1994).

Repeat number was found to vary within and between tissues, to be greater in muscle than in lymphocytes, and to be stable over 10 - 15 years, again suggesting postzygotic expansion. This was supported by the finding (Jansen *et al.*, 1994) of difference in repeat number between children's blood and fathers' sperm in paternal transmission. Muscle repeat number may correlate more closely with phenotype than does the usually-assayed blood repeat number. This *may* be an explanation for the finding that anticipation can occur even when the size of the allele (in blood) is reduced (Ashizawa *et al.*, 1994).

The coining of 'trinucleotide repeat disorders' and 'dynamic mutation'

In all of these three disorders the expansion was of a three-base repeat and they quickly became known as trinucleotide repeat disorders. Despite the fact that considerable study had already been made of the expansion of repeat sequences, these disorders were seen as having a new type of mutation. It was one that tended to predispose itself to further mutation and the probability of mutation of the product was different from the probability of mutation of the original sequence. Because of this it was quickly named 'dynamic mutation' (Richards & Sutherland, 1992a,b).

Further findings on SBMA

It was some months after the clear demonstration of the mechanism of anticipation in myotonic dystrophy that reports started to appear of correlation between disease severity and repeat number with the much shorter repeat tract of the

earlier-discovered SBMA locus (Doyu *et al.*, 1992; Igarashi *et al.*, 1992; La Spada *et al.*, 1992). In this disorder, unlike fragile-X syndrome and DM, the repeat tract was found to be more unstable on being transmitted by a male than by a female (La Spada *et al.*, 1992).

The position at the start of this project

1. Further reported inherited repeat instability disorders

a) Trinucleotide repeat disorders

Discoveries continued. By the time my research began, in May 1994, the position was as summarized in the reviews of Bates & Lehrach (1994) and Richards & Sutherland (1994). To *FRAXA* had been added another folate-sensitive fragile site on the X chromosome, *FRAXE*, also caused by expansion of a d(GGC)·d(GCC) repeat tract and found in families with mild mental retardation who did not have the expansion at *FRAXA*. In normal individuals there were no interruptions in the repeat tract and the longest allele observed was 25 copies. Affected males again began at about 200 copies and expressed the fragile site and again there was methylation of a nearby CpG island [later found to be associated with a gene, *FMR2* (Gecz *et al.*, 1996; Gu *et al.*, 1996)]. Carrier females with 116 - 133 repeats were unaffected and cytogenetically negative but those with >200 repeats expressed the fragile site and were possibly slightly mentally retarded. Here again, some males had been found to be mosaics, indicating expansion of the repeat tract in mitosis. The genetics had been found to be slightly different in that the repeat tract is equally unstable when passed on by a male as by a female, and contractions of the repeat are much more common than in the fragile-X syndrome.

Three other disorders had also been found to be caused by expansion of repeat sequences: Huntington disease (HD), spinocerebellar ataxia Type I (SCA1), and dentatorubral-pallidoluysian atrophy (DRPLA). These are all autosomal

dominant degenerative neurological disorders with variable severity and variable age of onset, showing anticipation, and in all three the mutation was found to be expansion of a d(CAG)·d(CTG) tract in the respective genes, coding for polyglutamine. In all three, as with SBMA, instability was found to be greater on transmission by males. In HD all juvenile onset cases were found to come from male transmission and in SCA1 and DRPLA some contractions were seen in female transmission (Bates & Lehrach, 1994; Koide *et al.*, 1994; Nagafuchi *et al.*, 1994).

As with SBMA, the range of repeat expansion was much smaller than seen in the disorders in which the repeat was non-coding, FRAXA, FRAXE and DM, a normal range of 6 - 39 and an abnormal range of 41 - 81 having been observed for SCA1 (Chung *et al.*, 1993; Matilla *et al.*, 1993) and 7 - 34 and 49 - 83 respectively for DRPLA (Koide *et al.*, 1994; Nagafuchi *et al.*, 1994). In Huntington disease the largest repeat length recorded was 121 and there was some confusion about where the dividing line lay, somewhere in the 30 - 38 region, between allele sizes that would or would not cause the development of symptoms. This was temporarily resolved by the discovery that there is a polymorphism in the length of a short d(CCG)·d(CGG) repeat closely following the d(CAG)·d(CTG) tract and included within the span of the PCR primers used by some laboratories (Rubinsztein *et al.*, 1993a,b) so that the length of the d(CAG)·d(CTG) tract was incorrectly estimated. However, this did not prove to be the whole answer. Though some affected individuals have repeats in the 36 - 39 range, others have been found with repeats in the same range who have remained symptomless beyond common life expectancy (Rubinsztein *et al.*, 1996; McNeil *et al.*, 1997).

Since the non-coding repeat whose expansion was responsible for DM was the same repeat that was involved in SBMA, HD, SCA1 and DRPLA, it seemed to us and to others (Zühlke *et al.*, 1993), that the repeat in these other disorders must also be capable of massive expansion and therefore that the reason that such large expansions were not seen was only that they were lethal at an early stage of development when they were coding, yet there are still those who interpret the small

expansion as indicating that position within a gene is a critical factor in determining the degree of instability of a sequence (La Spada, 1997).

b) Mismatch-repair deficiency

Another type of instability had also come to light. Eukaryotic genomes abound with repetitive DNA and repeat tract length polymorphisms form the basis of DNA fingerprinting and are invaluable as linkage markers for locating genes. To be useful for these purposes, a variable number tandem repeat (VNTR) must be unstable enough in the course of evolutionary time that it has produced many alleles so that nearly every individual is a heterozygote and most individuals are different at the locus, but it must be stable enough that a change in the number of repeats is hardly ever observed between one generation and the next or, of course, in mitotic cell division.

One class of cause of neoplasia is deletion of tumour suppressor genes. In 1993, three groups looking for genes responsible for non-polyposis colon cancer by using VNTRs to look for allelic loss in tumours compared with normal tissue found instead changes in repeat numbers (Aaltonen *et al.*, 1993; Thibodeau *et al.*, 1993; Ivonov *et al.*, 1993). These changes were mainly of one to four repeat units deletion or expansion and were found to be widespread in the genome and in mono-, di-, and tri-nucleotide repeats, including the ones that had been found to expand in trinucleotide repeat disorders. This was quite different from the situation in the trinucleotide repeat disorders in which only a single locus was involved in any particular disorder and changes in repeat number are often much larger.

Genes on chromosomes 2 (Peltomäki *et al.*, 1993) and 3 (Lindblom *et al.*, 1993) were found to be linked to this disorder and were then shown to be homologues of the mismatch-repair (MMR) genes *mutS* and *mutL,* respectively, of *E. coli,* and named *hMSH2* (Fishel *et al.*, 1993) and *hMLH1* (Bronner *et al.*, 1994). Subsequently two more mutant mismatch repair genes were identified in other cases of the disorder. They most resembled the yeast *mutL*-homologue *yPMS1* and so were

named *hPMS1* and *hPMS2* (Nicolaides *et al.*, 1994). (Throughout this thesis I shall use the term 'mismatch' loosely to include looping-out of a small number of bases.)

Change in the number of copies in tracts of short repeat units by a few units appears to occur by a mechanism variously known as 'strand-slippage', 'replication slippage', 'polymerase slippage' and 'slipped-strand mispairing', first proposed as a cause of mutation in DNA replication *in vivo* by Streisinger *et al.* (1966). In this, the growing tip of a nascent DNA strand is displaced between direct repeats on the parent strand, resulting in insertion or deletion of bases on the new strand depending upon the direction of the displacement. Displacement of one strand upon the other results in bases being 'looped-out' and if this is only a few bases it is normally detected and rectified by mismatch-repair enzymes. These detect loops of up to 4 bases in prokaryotes (Parker & Marinus, 1992) but up to at least 14 in eukaryotes (Fishel *et al.*, 1994).

2. The pattern of repeat location and expansion

Though there are 64 different codons, there are only ten different types of trinucleotide repeat. From the 64 trinucleotides one can first subtract the four that contain only one type of base (AAA *etc.*), as these do not repeat every three bases, and the other 60 have to be divided by two because there is another trinucleotide on the complementary strand, and further divided by three because, as regards purely the DNA sequence, rather than its translation, if any, the frame is unimportant. Thus the ten possible trinucleotide repeats are:

1. $d(AAC)_n \cdot d(GTT)_n = d(ACA)_n \cdot d(TGT)_n = d(CAA)_n \cdot d(TTG)_n$
2. $d(AAG)_n \cdot d(CTT)_n = d(AGA)_n \cdot d(TCT)_n = d(GAA)_n \cdot d(TTC)_n$
3. $d(AAT)_n \cdot d(ATT)_n = d(ATA)_n \cdot d(TAT)_n = d(TAA)_n \cdot d(TTA)_n$
4. $d(ACC)_n \cdot d(GGT)_n = d(CCA)_n \cdot d(TGG)_n = d(CAC)_n \cdot d(GTG)_n$
5. $d(ACG)_n \cdot d(CGT)_n = d(CGA)_n \cdot d(TCG)_n = d(GAC)_n \cdot d(GTC)_n$
6. $d(ACT)_n \cdot d(AGT)_n = d(CTA)_n \cdot d(TAG)_n = d(TAC)_n \cdot d(GTA)_n$

7. $d(AGC)_n \cdot d(GCT)_n = d(GCA)_n \cdot d(TGC)_n = \mathbf{d(CAG)_n \cdot d(CTG)_n}$

8. $d(AGG)_n \cdot d(CCT)_n = d(GGA)_n \cdot d(TCC)_n = d(GAG)_n \cdot d(CTC)_n$

9. $d(ATC)_n \cdot d(GAT)_n = d(TCA)_n \cdot d(TGA)_n = d(CAT)_n \cdot d(ATG)_n$

10. $\mathbf{d(CCG)_n \cdot d(CGG)_n} = d(CGC)_n \cdot d(GCG)_n = d(GCC)_n \cdot d(GGC)_n$

Of the seven trinucleotide repeat disease loci that had been identified by the beginning of this project, all contained one or other of only two of these repeat sequences (7 and 10 above). Two possibilities to account for this bias that might seem immediately obvious were either that other trinucleotide repeats also expanded but these two caused problems because they tended for particular reasons to occur in genes much more frequently than the others, or these sequences had some special property that made them more likely to expand than other sequences. However, there seemed to be much more interest in the second possibility, even after the first had been shown to be true.

a) Relative occurrence of trinucleotide repeat loci

First, it should be mentioned that two other trinucleotide repeats had been identified as occurring at highly variable loci with very high mutation rates and large size changes and were quoted as such by Jansen *et al.* (1994) in a paper on DM mutations. The first was $d[(CAC)\cdot(GTG)]_n$. Schäfer *et al.* (1988) reported that the probe $d(CAC)_5$ was very useful for human DNA fingerprinting. When human DNA was digested with *Hinf*I, subjected to electrophoresis and Southern blotting and hybridized with this probe, 252 bands were seen in the range 4 - 27 kb, of which an average of 15·8 bands were 'polymorphic per individual' (presumably they meant heterozygous). The bands of course related to many different loci.

Nürnberg *et al.* (1989) examined nuclear families with $d(CAC)_5$ or $d(GTG)_5$ probes and found that at one of these loci bands could undergo changes in length of 200 bp to 3 kb. However, they noted that the larger bands were not more intense as they should have been if they had been binding much more of the probe so they concluded that it was some other sequence that was unstable and the sequence

binding the probe was just on the same restriction fragment. Zischler *et al.* (1992) cloned and sequenced five of the loci binding the probe on four different chromosomes and found that one was $d[(ACC)\cdot(GGT)]_4$ and the others were all complex longer repeat sequences that contained some triplets that would bind the probe. Epplen & Epplen (1994) searched human lymphocyte cDNA libraries with the same probe and they found that stretches of ≤ 6 d(CAC)·d(GTG) trinucleotides were sometimes contained in open reading frames but more often in the 5' and 3' untranslated regions of mature mRNA and stretches of perfect simple d(CAC)·d(GTG) repeats longer than this were seldom recovered, even from hnRNA.

The second hypervariable trinucleotide repeat locus quoted by Jansen et al. (1994) was from the paper by Epplen et al. (1991) quoted in the first paragraph of this thesis and was found in the domestic hen with the probe $d(GAA)_6$. The authors noted that this locus was transmitted in a non-Mendelian manner. 80% of offspring had bands that were different from those of either parent. Sperm showed the same bands as somatic tissues and a pair of monozygotic twins had identical bands so they concluded that a rearrangement was occurring in early embryogenesis.

Homologous sequences were found in a wide variety of other vertebrates under intermediate to high stringency. The authors believed the sequence to be composed of a simple d(GAA)·d(TTC) repeat tract and pointed out that a long tract of this purine/pyrimidine asymmetry was bound to form secondary structure. They emphasized that no other locus ever found to that date had such a phenomenal instability, including other loci detected with the $d(GAA)_6$ probe in hens. They had cloned the unstable locus but not yet sequenced it.

The following month the fragile-X sequence came out (Verkerk *et al.*, 1991) and the nature of the hypervariable locus in the hen was still not published by 1994. However, the same authors did publish the results of a search of human DNA with their $d(GAA)_6$ probe (Siedlaczck *et al.*, 1993). They concluded that in human genomic DNA d(GAA)·d(TTC) repeats were underrepresented compared with other di-, tri- and tetranucleotide repeats and in foetal human brain cDNA there were very

few expressed d(GAA)·d(TTC) repeats. More recently I contacted Professor Epplen and the sequence in the domestic hen, in the particular allele cloned, was d[(GAA)·(TTC)]$_{24}$ (now puplished as Mäueler *et al.*, 1998).

The discovery of the first few trinucleotide repeat disorders engendered searches for other loci with the same repeats and loci with other trinucleotide repeat sequences. Some of these searches were by probing DNA and others by probing sequence-libraries. Searches that consisted of probing cDNA libraries with d(CAG)$_n$ and d(CGG)$_n$ probes were not useful for investigating the possibilities mentioned earlier. Numerous other genes containing these repeats were found and those which have since been associated with expansion disorders will be mentioned later.

In the most comprehensive of the sequence searches, Stallings (1994) searched the mouse, rat and human databases in GenBank (7·4, 5·0 and 16·4 Mb respectively at that time) for lengths of eight or more of all ten possible trinucleotide repeat sequences. In the human database, repeats of d(AAT)·d(ATT) and of d(TTG)·d(CAA) were about as common as those of d(CGG)·d(CCG) and d(CAG)·d(CTG), the numbers of occurrences being 10, 9, 9 and 12 respectively. Two were not found at all in any of the databases - d(GAC)·d(GTC) and d(TAC)·d(GTA) repeats - and d(CAC)·d(GTG) repeats were not found in the human database. The remaining three were found but were less common.

It has to be realized that the sequence databases would have been heavily weighted towards sequences containing genes so these frequencies would not necessarily represent the frequencies in the whole genome. One study relevant to this is that of Hummerich *et al.* (1994). They looked at the distribution of trinucleotide repeats across a 2 Mb cosmid contig known to contain the *HD* gene. They probed Southern blots of restriction digests of every cosmid with oligonucleotides consisting of six copies of every one of the ten types of repeat. They found 51 trinucleotide repeats including the one that was subsequently found (but sooner published) to be expanded in HD.

All ten classes of repeat were found. The commonest were d(TTG)·d(CAA) [probed with $d(CAA)_6$] and d(AAT)·d(ATT) [probed with $d(TAA)_6$] at 16 and 10 locations respectively. Both were also found to be common in Stallings' study. There were just four d(CAG)·d(CTG) occurrences and there were seven of d(CGG)·d(CCG). Of the three repeats that were not found in the human database by Stallings, $d(GTC)_6$ found three sites, $d(ACT)_6$ found one and $d(ACC)_6$ found two, suggesting that these sequences are also relatively uncommon in intergenic DNA in the human genome. They were not sequenced to see how long they were. That the statistical validity is fairly low is brought out by the consideration that two sites binding $d(ACC)_6$ in 2 Mb represents 1 site/Mb and Schäfer *et al.* (1988) found 252 fragments binding $d(CAC)_5$ in the whole 3,000 Mb of the human genome and some of these were allelic.

Returning to Stallings' (1994) database search for strings of at least 8 perfect repeats, to investigate why d(CAG)·d(CTG) repeats were the most prevalent in human trinucleotide repeat expansion disorders he then sorted all those sequences sufficiently annotated (95 from all 3 species combined) into 5' untranslated, exon, intron and 3' untranslated locations. For $d[(CGG)\cdot(CCG)]_8$ they were 3, 5, 0 and 0, and for $d[(CAG)\cdot(CTG)]_8$ the numbers in these categories were 1, 14, 0 and 0, *i.e.* mainly in exons and none in introns. (For d(CAG)·d(CTG) this was strand-specific for $d(CAG)_8$ on the coding strand as the respective numbers for $d(CTG)_8$ were 3, 4, 3 and 1.) Conversely, the other two most common trinucleotide repeats were almost absent from exons, d(AAT)·d(ATT) occurring as 2, 0, 11 and 1 in the same categories respectively and d(TTG)·(CAA) with 2, 1, 5 and 1. Stallings suggested that the preponderance of d(AAT)·d(ATT) in introns might be related to a regulatory rôle. (He found just two d(GAA) repeat sites on the coding strand and they were both in introns; d(TTC) repeats were also found in two introns and two 5' untranslated regions.)

For d(CAG)·d(CTG) and d(CGG)·d(CCG) the search was then extended to all runs of at least 4 repeat units to see whether any would fall into introns. For

d(CGG)·d(CCG) a very few did, the numbers now coming to 59, 75, 4 and 2 for the four categories, but of the $d(CAG)_4$ tracts on the coding strand 12 were in 5' untranslated regions and 96 in exons with still none in introns or 3' untranslated regions. Finally Stallings searched for $d(CAG)_8$ with one or two imperfections and ten more sequences were found and none of them were in introns either. He suggested that selection might exclude d(CAG) repeats from introns because of their similarity to the 3' acceptor consensus splice site which contains the highly conserved sequence CAGG.

For any gene with a trinucleotide repeat in one of the three species a search was made for a sequenced homologue in the other two. Comparison of these genes showed little conservation of repeat tracts; a long tract in one species might be small absent or in a different position in the homologous gene of another species. The same applied when comparison was made of amino-acid repeats coded by trinucleotide repeats. Seven of the genes found in this survey (only one of them known to be associated with an expansion disorder) contained two repeat tracts. Some of these were very close together in the gene and/or differed by only one base, *e.g.* d(ATT) and d(TTG) repeats in the 5' flanking region of the rat α_2-macroglobulin gene and d(GAG) and d(GAT) in the coding sequence of the human histidine-rich calcium binding protein gene.

Green & Wang (1994) looked at codon reiteration in proteins in all species covered by three databases. For runs of 5 - 9 identical amino-acids the frequency of the tract was roughly proportional to the frequency of the amino-acid, with a few exceptions that were hardly if ever in runs even as short as this. Leucine, the most common amino-acid had the most runs and glutamine, the fifteenth in amino-acid frequency was seventh in the order of runs. When the length of run was increased to 10 - 14, leucine became uncommon. Glutamine's frequency had dramatically increased to be the most common but alanine, glycine, serine glutamic acid, asparagine and histidine runs were still quite common in that order. As the length was increased,

these dropped off too, leaving glutamine outstandingly dominant in tracts of ≥20 residues.

Of the 229 reiterants of ≥10 amino-acid residues, only 48 were encoded by uninterrupted repeats of a single codon. Of these, 21 were encoded by the three codons that are CAG or a circular permutation of it, 16 were encoded by members of the 16 codons that contain CA, AG or GC in any frame, and only 11 were encoded by members of the other 42 coding codons.

b) Expansion of other trinucleotide repeats

Schalling *et al.* (1993) invented a method of detecting expanded repeat arrays, which they called RED for repeat expansion detection. It uses a single fairly long detecting-oligonucleotide, *e.g.* $d(CTG)_{17}$, a thermostable DNA ligase and a PCR machine and hence was dubbed the ligase chain reaction, but it is not such. It cycles through denaturation, annealing and ligation. If a repeat tract complementary to the oligonucleotide is short, *e.g.* 4 repeats, the chances of two oligonucleotides annealing to it *in tandem* are remote but if the array is expanded it can accommodate several or very many oligonucleotides and if two anneal adjacent to one-another they will be ligated. The production of concatemerised oligonucleotides is linear, not exponential. After several hundred cycles, if and only if there is an expanded array in the template, electrophoresis will display a ladder of 1, 2, 3, -n oligonucleotides ligated together.

By this method, the authors discovered (Lindblad *et al.*, 1994) arrays of 60 or more units of four repeats apart from the two that had so far appeared at disease loci: d(ATG)·d(CAT), d(CCT)·d(AGG), d(CTT)·d(AAG) and d(TGG)·d(CCA) (lines 9, 8, 2 and 4 on pp. 16-17). It may be noted that these repeats include the two repeats that had been detected at hypervariable loci before the dawning of the trinucleotide repeat age by Nürnberg *et al.* (1989) and Epplen *et al.* (1991).

In summary then, the two repeats that had so far been found in expansion disorders were not the only trinucleotides to occur in polymorphic arrays nor the

ones to expand. Their association with disease did seem to be because they are the most strongly associated with exons and 5' untranslated regions of genes. Any explanation of trinucleotide repeat expansion must therefore take all the expanding sequences into account.

3. Other highly unstable repeat sequences

Jeffreys *et al.* (1987) noted than one of a number of mouse DNA fingerprint loci that they identified with human minisatellite probes had a very wide variation in allele length and a very high mutation rate and named it *Ms6-hm*. By linkage analysis they found and cloned the locus (Kelly *et al.*, 1989). The repeating element was d(CAGGG)·d(CCCTG) [or circular permutation thereof; the sequence started with GGGG and ended with GGG and by comparison with similar sequences may have evolved from a progenitor that was a single d(GCAGG)·d(CCTGC)]. The length of the cloned DNA fragment collapsed in *E. coli* in stages from 7 kb down to 334 bp, losing almost all of its estimated 1,340 repeats down to 19 in the sequenced remnant. The germline mutation rate in mice was found to be about 2·5% per gamete and occurred in the paternal allele in 8/8 cases investigated. Mosaicism was observed in 3% of mice and pointed to somatic mutation in very early development.

A $(GGGCA)_n$ probe from the *Ms6-hm* locus hybridized to other mouse loci. Many of these also appeared to be highly unstable and at one of these, *Hm-2*, somatic mutation was also observed. This locus was subsequently also cloned (Gibbs *et al.*, 1993) and found to have the repeat unit d(GGCA)·d(TGCC) [or d(GCAG)·d(CTGC)] with alleles containing from ~500 to ~6,500 copies. The germline mutation rate was estimated at 3·5% (95% limits 2·3 - 5·1%) but was considered to be an underestimate as a change in length of ~50 repeats was the lower limit of their resolution and there were probably many smaller changes. The change in repeat number in one generation varied up to ~2,200 with no significant bias to gain or loss or to parental origin.

Mosaicism was seen in 20% of adult mice with no bias of sex or parental origin but in one particular cross of strains there was an excess of somatic mutations with a significant bias to decrease in size and with significantly more mutations in the F_1 than the F_2 generation. Detailed analysis of embryos, trophoblasts and yolk-sacs revealed that most somatic mutations occurred during the first two cell divisions and rapidly decreased thereafter with possibly all occurring before the fifth division (32-cell stage). The authors speculated that this may reflect the transient expression of a factor involved in repeat instability or possibly carry-over of a meiotic factor into the zygote. Candidate proteins were the minisatellite binding proteins which had been reported by several authors.

The same laboratory also identified several very unstable human satellite loci. Jeffreys *et al.* (1988) investigated five minisatellite loci in 40 large families and one of these, MS1 at locus *D1S7*, which is a 9 bp repeat, d(GTGGA^{c}/tAGG)· d(CCT^{a}/gTCCAC) (Wong *et al.*, 1987) was found to have a mutation rate of 5·2% per gamete within the resolution of testing. The number of repeats ranges from ~50 (Jeffreys *et al.*, 1988) up to ~3,000 (Gibbs *et al.*, 1993) and the largest change observed by Jeffreys *et al.* (1988) was ~200 repeat units. The lower limit of resolution of their gels was about 4 repeat units and the authors estimated that the true mutation rate might be about 10% per gamete. The rates of gain and loss were similar, indicating no substantial bias towards expansion, suggesting that the largest alleles accumulated fortuitously by genetic drift.

Other loci with fairly high mutation rates reported by the same laboratory were MS31, a 20 bp repeat, d(TGGGAGGTGG^{a}/g^{c}/tAGTGTCTG)· d(CAGACAC T^{a}/g^{c}/tCCACCTCCCA)(Wong *et al.*, 1987) having a mutation rate of about 0·7% (Jeffreys *et al.*, 1988) and MS32, a 29 bp repeat, d(GAATGGAGCAGG^{c}/tG^{a}/gCC AGGGGTGACTCA)·d(TGAGTCACCCCTGG^{c}/tC^{a}/gCCTGCTCCATTC) (Wong *et al.*, 1987) of length 12 - 800 repeats (Jeffreys *et al.*, 1994) and a mutation rate of about 1·2% per gamete (Jeffreys *et al.*, 1991). Another laboratory (Vergnaud *et al.*, 1991) found a minisatellite, CEB1, with a much higher mutation rate, but no instance

of mosaicism was found. 75 mutations were identified amongst 565 children from 61 families and in only two instances was the mutation of maternal origin, giving a mutation rate of about 13% in spermatogenesis against only about 0·4% in oogenesis (Buard & Vergnaud, 1994). Eight different variable positions were found amongst 20 copies of the repeat unit sequenced from one allele. The repeat units varied in length from 37 - 43 bp. One strand was very C-rich (and the other correspondingly G-rich). Allele sizes ranged from <500 bp - >12 kb (Vergnaud *et al.*, 1991), *i.e.* <13 - >400 repeat units.

Mahtani & Willard (1993) reported that a human tetranucleotide repeat on the X chromosome, $d(TATC)_n \cdot d(GATA)_n$ where n = 12 - 16, had a high mutation rate. They found 4 mutations in 274 parent-to-child transmissions. This is higher than has been recorded for the disease-associated trinucleotide repeat tracts when they are that short.

Another human repeat locus, causing disease on expansion, had been reported before any of the 'trinucleotide repeat disorders' but seems to have escaped notice of most workers for several years. In 1990, Fearon *et al.* published the results of their search for a tumour supressor gene on chromosome 18q. Allelic deletions in this region had been found in 70% of colorectal cancers. Fearon *et al.* (1990) found that expression of the gene *DCC* (for Deleted in Colon Cancer) was greatly reduced or absent in colorectal carcinomas and that the gene had somatic mutations that were not present in normal tissue. One type of mutation that was seen in ten cases was an apparent insertion. Normal alleles were found to vary in length in this region. The difference in size between the largest and smallest of 88 normal alleles was ~50 bp but the largest of these was about 120 bp smaller than the smallest of the altered alleles.

The corresponding region of one of the normal alleles was sequenced. It was found to contain two tracts of perfect d(TA)·d(TA) repeats, one of 8 and one of 26 copies, both within a 130-bp region of alternating purine and pyrimidine base pairs (mainly d(TG)·d(CA) with d(TA)·d(TA) and d(CA)·d(TG) interruptions) that could

potentially form Z-DNA. Fearon *et al.* (1990) were unable to clone alleles with intact copies of the expansion from any of the tumours tested; deletions arose during the cloning process in both bacteriophage and plasmid vectors in several *E. coli* strains, including one from this laboratory that allows cloning of repeated sequences that are difficult to clone in other hosts). The authors were also unable to amplify the enlarged alleles by PCR though they could easily amplify the normal alleles. This suggested that the enlarged alleles had an unusual structure that might also interfere with transcription. It seems likely that the enlargement was expansion of a repeat tract, probably of d(TA)·d(TA). It has not been recorded whether this expansion has been responsible for any familial cases of colon cancer. Had it been so, the term 'trinucleotide repeat disorders' might have been out of date even before it was coined. Certainly it is out of date now, as will be seen in a later section.

4. Trinucleotide repeat binding proteins

Richards *et al.* (1993) were interested in the fact that the repeat sequences responsible for the fragile-X syndrome and DM were in untranslated regions of their respective genes. They noted that it had been shown that the one in *FMR1* had been conserved in position throughout mammalian evolution though the distribution and number of repeats varied. They therefore deduced that some function other than coding was affected by expansion of the repeats. They also noted that d(CGG)·d(CCG) repeats had been found in the 3'-untranslated regions of several other genes and that in one of them, the *BCR* gene, the repeat sequence had been identified as one of eleven binding-sites for nuclear proteins 5' to the gene by DNaseI footprinting with nuclear extracts from human cell lines (Zhu *et al.*, 1990). Richards *et al.* (1993) therefore searched for nuclear binding proteins for simple tandem repeat sequences.

Complementary oligonucleotides were annealed to make 30 bp double-stranded DNAs of all ten possible trinucleotide repeats and all four possible dinucleotide repeats [d(AC)·d(GT), d(AG)·d(CT), d(AT)·d(AT), and d(CG)·d(CG)].

Each of the sequences was found to bind one or more nuclear proteins (from HeLa cells) with some degree of specificity. The one binding d(CGG)·d(CCG) repeats did not bind any of the other repeats or the consensus binding sites for Sp1 and AP2, which are similar, and was named CCG-BP1. 6 seemed to be the minimum number of repeats to which it would bind. If the sequence was methylated, binding was very much reduced and another protein bound. d(AG)·d(CT), d(AAG)·d(CTT), and d(AGG)·d(CCT), repeats all competed for the same protein and the authors commented that these were all polypurine/poly-pyrimidine sequences and that since such sequences were known to be able to form non-B DNA conformations the protein might be recognizing a particular structure rather than a specific sequence. The other repeats all bound unique proteins though for d(CAG)·d(CTG) and d(GAC)·d(GTC) repeats the proteins were either of very low abundance or low affinity under the conditions tested.

Because of the possibility of single strands of repeat sequences forming secondary structure, the single strands were also tested for specific protein binding. When labelled $d(CCG)_{10}$ was incubated with nuclear extract and subjected to electrophoresis, several bands were seen, one of which had the mobility of CCG-BP1. $d(CGG)_{10}$ bound a single major protein of a different mobility. Several other single strands of repeats also bound apparently distinct proteins. The authors suggested that binding proteins which might interfere with replication or stabilize alternative conformations of the repeat might facilitate slippage-mediated amplification at the replication forks.

5. Observations on trinucleotide repeat alleles

a) Uneven distributions of allele sizes

Observations of the distributions of different alleles *(i.e.* histograms showing proportion of alleles that are a given length plotted against length) revealed that normal alleles at these loci tended to have a bimodal or trimodal distribution. This

suggested that there might be two or three mechanisms at work. At the DM locus it was found that 35 - 40% of normal alleles had 5 repeat units, 50% had 11 - 14 and about 10% had 19 - 30 units (and there were a few of other lengths). This alone would suggest that there had been very rare large jumps of, say, 8 - 20 units at once, and then more common replication slippage events in either direction from these positions (when mismatch-repair had an occasional failure, or perhaps when the gene was passing through an individual who had a mismatch-repair defect). This was supported by the finding of linkage disequilibrium. Disease-causing alleles at the DM locus were found to be in complete linkage disequilibrium with an adjacent insertion/deletion polymorphism and the same allele of this polymorphism was also in complete linkage disequilibrium with the normal alleles of 5 and of 19 - 30 repeat units (Imbert *et al.*, 1993). This suggested that there might have been one, or at most a very few, expansions from five copies into the 19 - 30 range and that alleles in this upper end of the normal range provide the pool from which come further expansions into the pathogenic range. Subsequent studies (Rubinsztein *et al.*, 1994; Zerylnick *et al.*, 1995) showed that the disequilibria were not as strong as had been supposed and that there had probably been several initial expansion events.

Within the pathogenic range it was seen that there are sometimes very much larger expansions of hundreds of repeats or more. Sometimes the new allele in FRAXA or DM may be up to ten times the length of the parent allele from which it is found to have expanded (Kuhl & Caskey, 1993), which suggested that there must have been a repetitive expansion process within a single cycle of replication, or within a few cycles in the generation of a gamete. It was predicted by Laird (1987) that replication of *FRAXA* might be delayed and this was shown to be the case by Hansen *et al.* (1993). However, the latter pointed out that this delay could not be explained solely by stalling at the expanded repeat of a replication fork travelling from a proximal or distal origin because replication was delayed on both sides of the repeat up to a distance of at least 150 kb one way and at least 34 kb the other. They mentioned the possibility that the expanded repeat might itself act as an origin of

replication but this was later rendered unlikely when it was shown that the FRAXA repeat expands in a polar manner, *i.e.*, nearly all expansion occurs at just one end (Kunst & Warren, 1994).

b) Repeat tract interruptions

Another type of observation of normal and pathogenic alleles that had been made was of the patterns of repeat interruptions. Richards & Sutherland (1992b) had suggested that loss of the d(AGG)·d(CCT) interruptions in the fragile-X repeat tract, making longer stretches of perfect d(CGG)·d(CCG) repeats, might create the condition for instability, and it had since been shown that loss of interruptions did appear to operate in this way in SCA1. In SCA1, the d(CAG)·d(CTG) repeat tracts in normal individuals were nearly all found to be interrupted by one or more d(CAT)·d(ATG) trinucleotides whereas the tracts in pathologically expanded alleles were perfect d(CAG)·d(CTG) repeats and those normal alleles which had no interruptions had tracts at the lower end of the normal range (19 or 21 copies) (Chung *et al.*, 1993). Soon, loss of interruptions was also shown to be a predisposing factor in the FRAXA repeat too (Kunst & Warren, 1994; Snow *et al.*, 1994; Hirst *et al.*, 1994).

6. Evidence and hypotheses for the mechanism(s) of expansion

There was little if any question that small changes of one or two repeat units in either direction occurred by strand-slippage in replication. The questions were of what caused the larger changes to occur in a single generation and why in these mutations expansion so heavily outweighed contraction in most cases. It might even be wondered whether there might be one mechanism responsible for changes by tens of repeat units and another causing changes of many hundreds of repeat units. All of the proposed mechanisms fell into two classes - strand-slippage and recombination - or a combination of the two. This section is intended to relate what discoveries and

suggestions had already been made on the mechanism(s) of repeat tract expansion before the start of this project. Some reference will be made to more recent papers that have developed these ideas but their details will be left till later.

a) The development of ideas - loops, slippage and recombination

Most if not all the mechanisms had already been suggested or even demonstrated to occur in repeat tracts long before the discovery of trinucleotide repeat disorders. Fresco & Alberts (1960) investigated the mechanism by which mutation by deletion or misincorporation of bases during replication could occur. They synthesized partially complementary RNA strands (with polynucleotide phosphorylase) and hybridized them and found that double helices could form with unpaired bases looped out and showed by models that the same might happen with DNA. From this they deduced that substitution, deletion or addition of one or more bases during DNA replication could occur without halting the growth of the new helix. They also suggested that a helix with loops might provide a substrate for recombination.

Chamberlin & Berg (1962) purified an RNA polymerase from *E. coli* and found that if it was provided with rATP in the absence of the other ribonucleotide triphosphates, poly-rA was synthesized. This was dependent upon the presence of a DNA template even though the template did not contain a sufficiently long run of thymidines. They introduced the term "slippage" for the process by which they deduced that synthesis must occur by successive melting and reannealing to an oligo-T stretch of the DNA. This was supported by the work of Falaschi *et al.* (1963) who found that poly-rA was synthesized *in vitro* from a template of dT_5.

Kornberg *et al.* (1964) showed that high-molecular-weight poly-AT DNA could be synthesized by DNA polymerase starting from $d[T(AT)_n]$ oligomers. Longer primers had higher temperature optima for initiating the process, implying that repeated melting and reannealing at a new, displaced position must be occurring. The authors pointed out that an important possibility in this process might be of

hairpin structures forming in the new strand. They also suggested that slippage with looping-out of bases on the new strand could account for poly-d(ATA)·d(TAT) stretches in natural DNA as had been observed in a crab.

Extension of the work to the synthesis of long double-stranded repeat tracts of trinucleotides (Wells *et al.*, 1967a) and tetranucleotides (Wells *et al.*, 1967b) *in vitro* was then also recorded. The trinucleotide templates included $d(TTC)_n$, $d(GAA)_n$, $d(AAG)_n$, $d(TTG)_n$, $d(AAC)_n$, $d(TAC)_n$, $d(TAG)_n$, $d(ATC)_n$ and $d(ATG)_n$ with values of n up to 7. The authors (Wells *et al.*, 1967b) found that the rates of growth of the new DNA strands decreased considerably when the length of the repeating unit was increased from 2 to 3 to 4 for a 'readily understandable' reason[2].

In the meantime Streisinger *et al.* (1966) proposed slippage during DNA replication of repeat tracts *in vivo* to account for some of the frame-shift mutations observed in the lysozyme gene of bacteriophage T4. They mentioned that the frequency of this mutation would be likely to increase with the length of the repeat tract and found that at one site mutation from 6 to 5 adenine residues was at least 100 times as common as mutation from 5 to 6. Other reported mutations agreeing with his slippage model were a change from 3 to 2 or 4 repeats of d(CTGG)·d(CCAG) in the *E. coli lacI* gene (Farabaugh *et al.*, 1978) and other frame-shift mutations in the lysozyme gene of bacteriophage T4 caused by length change in repeats of single bases (Pribnow *et al.*, 1981). Thus not only length variations in non-coding satellite DNA but mutations in genes had been reported to be due to changes in numbers of repeats

---

[2] Levinson & Gutman (1987) later pointed out that this was evidence for the expectation that slipped-strand mutation *in vivo* should have an appreciable bias towards expansion of *short* repeat units. Some investigators have found higher rates of mutation in tetranucleotide repeats than in dinucleotide repeats (e.g. Weber & Wong, 1993 investigating 28 short tandem repeat loci in 40 CEPH families). However, Chakraborty *et al.* (1997) have subsequently concluded that dinucleotide repeats do have higher mutation rates than tetranucleotide repeats, with non-disease-causing trinucleotide repeats coming in between and disease-causing trinucleotide repeats exceeding all the others. They suggest that the reports of tetranucleotide repeat mutation exceeding that of dinucleotide repeats might stem from non-random sampling of tetranucleotide loci in direct mutation assays.

long before the discovery of trinucleotide repeat disorders but these were changes of a single repeat unit, not huge expansions.

Slightom *et al.* (1980), in the first report of gene conversion, found that a region of simple repeat sequences [d(TG)·d(CA) and d(CG)·d(CG) repeats in an intron] appeared to be a hot-spot for initiating the recombination. Hentschel (1982) showed that short tandem repeats are sensitive to S_1 nuclease, providing direct evidence that such sequences are prone to form single-stranded regions. The repeats were $d[(GA)\cdot(TC)]_{16}$ and $d[(CA)_{10}(CT)_{22}\cdot(AG)_{22}(TG)_{10}]$ in spacers between sea-urchin histone genes. Hentschel's results suggested that there could be slippage between repeats in unbroken double-stranded DNA making a loop at one end of the slippage on one strand and a loop at the other end of the slippage on the other strand and he speculated that this could occur *in vivo* and act as a focus for recombination.

Ripley (1982), inspired by the suggestion of Streisinger *et al.* (1966) that frame-shift mutations might be caused by slippage, sought to explain frame-shifts that could not be accounted for by slippage of direct repeats. She suggested that quasi-palindromic sequences might form hairpin or cruciform structures and that mismatches in the stems of these hairpins, which might be of 1 - 4 bases looped out from the stem, might be repaired, using either of the sides of the same stem as a template, resulting in insertions or deletions. Though she did not mention it as such, this would tend to perfect palindromic sequences. Ripley (1982) found that this and another possible mechanism involving quasipalindromes could explain more than 15% of frame-shift mutations in the iso-1-cytochrome *c* gene of *Saccharomyces cerevisiae* and subsequently (de Boer & Ripley, 1984) some base substitutions too. This type of mutation certainly does not seem to occur in trinucleotide repeat tracts. There have been no reports for instance of mutations in a d(CTG) strand converting triplets in one half of the tract to d(CAG). However, the suggestion of quasi-palindromic sequences forming secondary structures that led to mutation was important as will be seen.

In 1992, only the first three trinucleotide repeat disorders had been discovered. These involved only two types of repeat and both were of the form d(CXG)·d(CX'G), where X and X' are complementary bases and Sinden & Wells (1992) were the first to suggest that these sequences, due to their quasipalindromic nature, might form hairpins held together by C·G bonds or, in the case of $d(CGG)_n$ single strands, possibly by G·G bonds, with the other bases mispaired, and that these hairpins might aid the slippage process just as had been suggested for the perfectly palindromic $d(AT)_n$ strands by Kornberg *et al.* (1964). They did not say that these sequences might be the ones found in inherited disorders because they might be more prone to expand than others because of the possible hairpin-forming ability of their single strands, but this seemed increasingly to be the view (Smith *et al.*, 1994; Mitas *et al.*, 1995a; Gacy *et al.*, 1995; Chen *et al.*, 1995; Smith *et al.*, 1995) as more disorders were added with the same repeat sequences. Sinden & Wells (1992) did not even propose that this possible hairpin formation might be the main cause of the slippage. They envisaged that there might be a more stable secondary structure, perhaps a triplex or a quadruplex, blocking the progress of the replication fork, and that this might cause the polymerase to slip back and readvance repeatedly on the leading strand, building up an ever-lengthening hairpin of the new strand behind it, though it was later pointed out (Mitas *et al.*, 1995a) that this model went against evidence that suggested that mutation usually occurs on the lagging strand.

For expansion to predominate over contraction, a model of hairpin-enhanced slippage requires that secondary structure formed by the nascent strand be more stable than that formed by the template strand. Though it is likely that one strand might form more stable secondary structure than the other, because of their different base compositions [*e.g.* $d(CAG)_n$ and $d(CTG)_n$], one has to remember that on the other arm of the replication fork the positions are reversed as to which base sequence is the template and which the new strand. Thus, unless there is to be a contraction of the repeat tract on the other arm of the fork, one has to postulate that secondary structure only forms on one arm, or only causes a problem on one arm.

Returning to Ripley (1982), the other mechanism by which she proposed that quasipalindromic sequences might cause frame-shift mutations involved hairpin formation after strand switching in replication. Her drawing is reproduced in Figure 1.1.

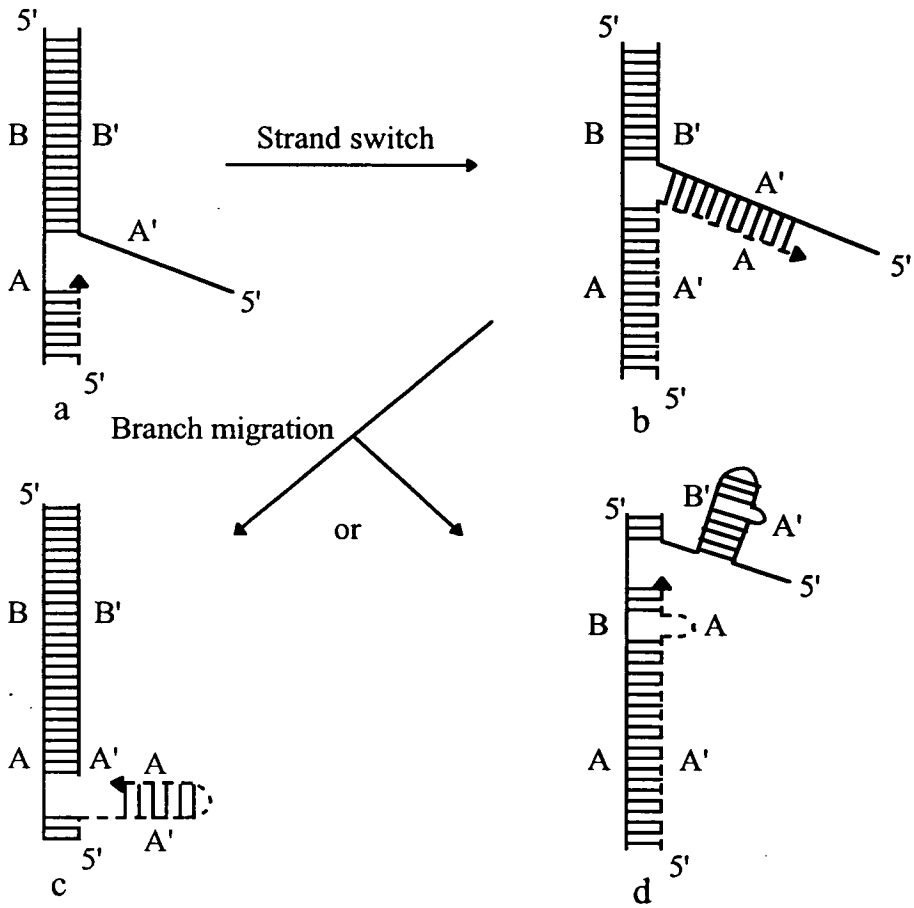


Figure 1.1 One of the models of Ripley (1982) for how quasipalindromic sequences might cause frame-shift mutations.

Ripley's drawing was intended to represent a replication fork. She envisaged that the hairpin formed in resolution (c) would be removed and that no mutation would result whereas in resolution (d) repair of the mismatch of B with A and of B' with A' would lead to frame-shift. If instead we take the drawing as being of extension of an Okazaki fragment on the lagging strand of a replication fork, pushing out a flap on the 5' end of the fragment in front of it, and the sequence d(BA)·d(A'B') not as some non-

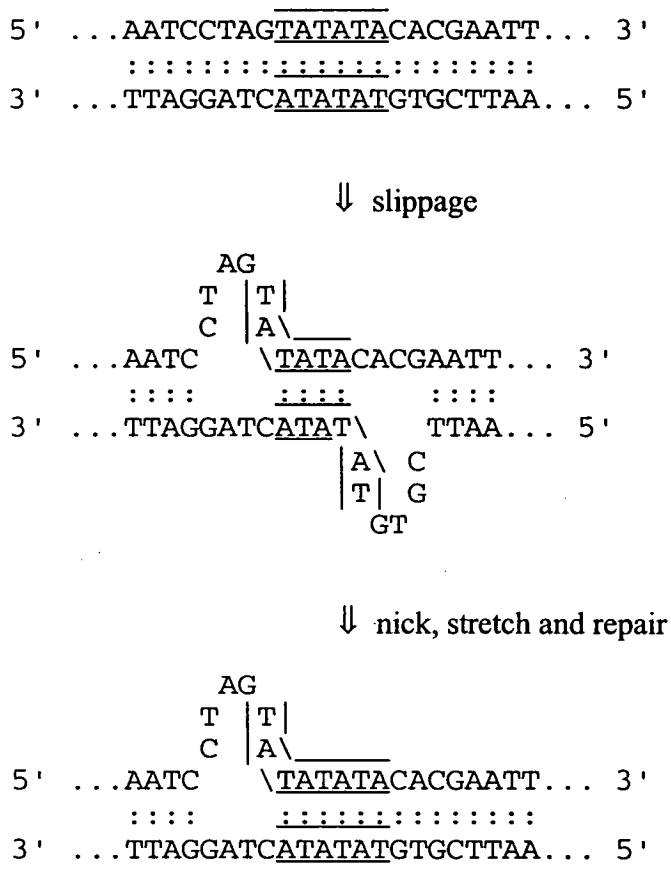
repeating sequence that happens to be quasipalindromic but as a d(CXG)·d(CX'G) repeat tract, and if we go directly from (a) to (d) without the strand switch (and therefore do not make a mismatch of A with B but still have the B' to A' mismatch) then we can see that if the hairpin in (d) were not cleaved off, expansion could result either by ligation of the free ends and further replication or by recombination with the other arm of the replication fork. (The B' to A' mismatch would represent, for example, an A·A mismatch in a hairpin of d(CAG) repeats.) This very model of expansion has subsequently been proposed by Gordenin *et al,* (1997).

In a general study of simple repetitive DNA sequences, Levinson & Gutman (1987) concluded that replication-slippage is likely to be the major mechanism in initial expansion of short tracts of repeats and that after the initial expansion the sequences might be predisposed to further expansion by unequal crossing-over or other interhelical events, such as gene conversion, because of their propensity to mispair. They pointed out that both mechanisms would be likely to have a self-accelerating component (*i.e.* dynamic mutation by another name, and well before 1992).

Levinson & Gutman (1987) also discussed the observation that in prokaryotes mutations in repeated sequences were skewed towards deletions whereas in multicellular eukaryotes there was clearly a bias towards expansion so here again, the much-discussed observation that the repeat sequences that were involved in human inherited disorders expand much more often than they contract was not a new one. Levinson & Gutman (1987) suggested two main possibilities. One was that bacteria might have evolved a bias towards deletions in order to minimize genome size because of selective pressure for rapid replication and that genome size might be less important in multicellular eukaryotes whose genetic apparatus might tend to favour insertions over deletions either by generating more insertions or by repairing insertion heteroduplexes less efficiently. The other was that there might be selective pressures that tend to favour long-term retention of simple-repeat DNA, though there would have to be a system to keep this under

control, and that selection against accumulation would of course be expected to the major factor in coding sequences.

Taking up Hentschel's (1982) suggestion that there could be slippage between repeats in unbroken duplex DNA making a loop at one end of the slippage on one strand and a loop at the other end on the other strand, Levinson & Gutman (1987) suggested that this could lead to expansion if there was nicking and repair of one strand opposite the loop. Their little diagram (2B p. 207) is redrawn below.



Recently this very same mechanism has been proposed for expansion of trinucleotide repeats by Petruska *et al,* (1998). The differences were only that they had more repeat units and proposed that a cruciform would first form in the repeat DNA with each strand self-annealing to form hairpins; then the hairpins would migrate apart, each moving in a 5' direction, before nicking and repair as before. Levinson & Gutman (1987) did even mention that branch migration of single-stranded loops was

likely. The trouble with these suggestions is surely the question of why one of the structures would be repaired and the other not. (If the top strand in the diagram was being repaired then why would not the loop on that strand be removed and the sequence then restored back to the original?)

More complex models involving strand-slippage or recombination were also put forward. DNA polymerase was found to have difficulty in replicating the repetitive CG-rich sequence in the folate-sensitive fragile sites (Yu *et al.*, 1991; Kremer *et al.*, 1991; Fu *et al.*, 1991; Knight *et al.*, 1993) and it was suggested (Fu *et al.*, 1991; Kuhl & Caskey, 1993) that repeated stalling and reinitiation of replication might produce an 'onion skin' structure of replication bubbles within bubbles, as had been proposed for gene amplification (Stark & Wahl, 1984). Unequal recombination between these strands could then produce the repeat expansion (Fu *et al.*, 1991; Stark & Wahl, 1984). Alternatively, expansion could also be produced by switching of a growing strand from one arm of the onion skin to another (Kuhl & Caskey, 1993), another variation of strand slippage. In either case, the replication stalling might be brought about by the formation of secondary structure. One other suggestion (Richards & Sutherland, 1994) is better discussed later.

Examples of ways in which secondary structure in repeat DNA might be involved in dynamic mutation by strand-slippage or recombination during replication are illustrated in Figure 1.2 (overleaf). Slippage is represented on the lagging strand of the replication fork since several studies suggest that, for hairpins formed by palindromic DNA, slippage occurs more frequently there than on the leading strand (Leach, 1994). In the recombination model, resolution is shown by disengagement of strands as proposed in several double-strand break repair models (Resnick, 1976; Nasmyth, 1982; Thaler *et al.*, 1987; Hastings, 1988). If recombination with a homologous chromosome is possible, this is necessary to explain the lack of crossing-over associated with repeat expansion (Imbert *et al.*, 1993; Kunst & Warren, 1994; Jeffreys *et al.*, 1994) [see discussion below under (c)]. Resolution by cleavage of the

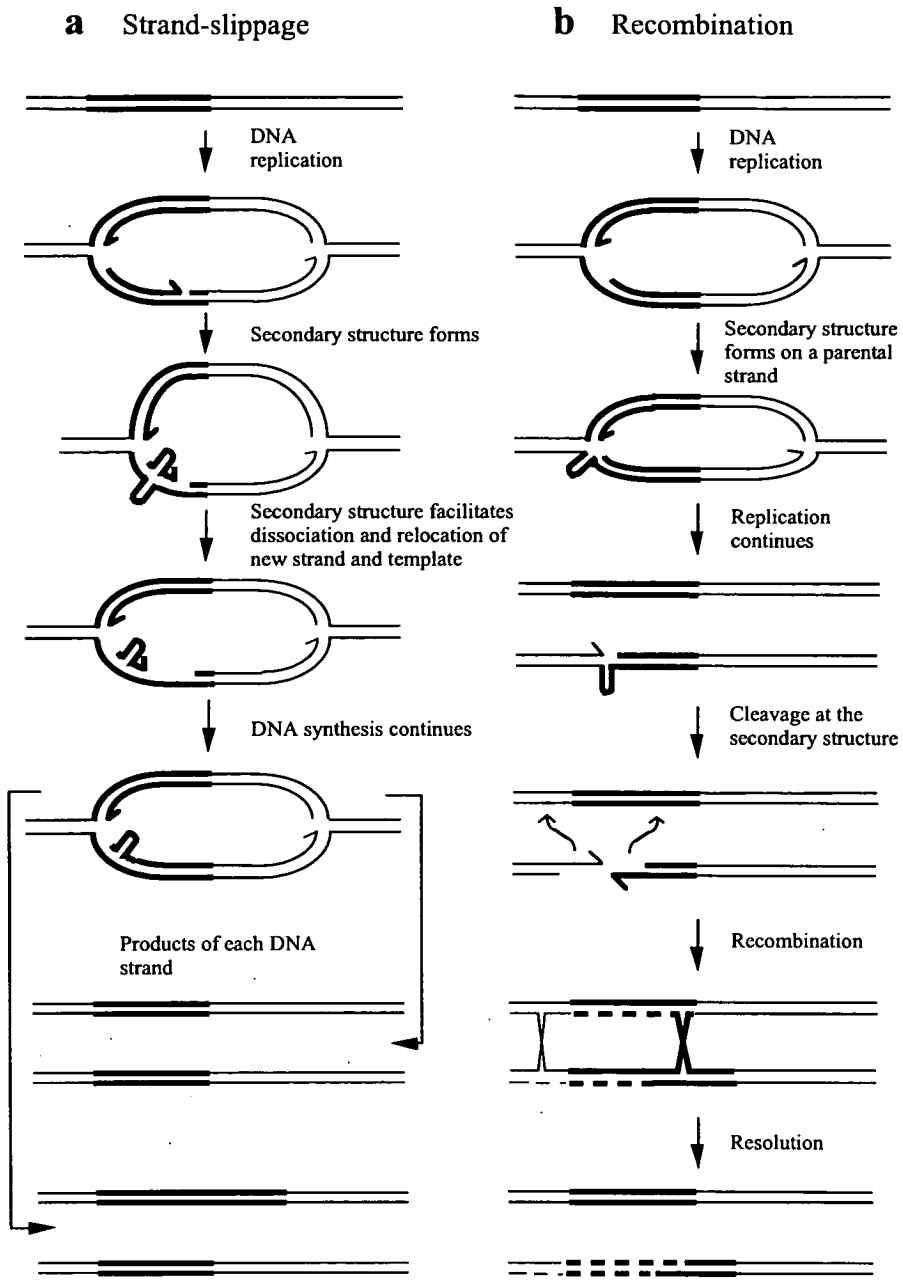


Figure 1.2. Two classes of model for dynamic mutation and the potential involvement of unusual DNA secondary structures in each. The tract of trinucleotide repeats is represented by thick lines and normal DNA by thin lines. The secondary structure is represented by a hairpin but may be more complex. (a) Replication-slippage can account for amplification if the newly synthesised strand folds back on itself and then replication re-copies the section of template previously used. (b) Recombination is known to be stimulated at sites of double-strand breaks. In a tract of trinucleotide repeats, the formation of a secondary structure may lead to cleavage of one sister chromatid and its repair by recombination. The broken arm may recombine at many sites within the repeated array to generate new alleles of variable length. (Reproduced from Darlow & Leach (1995), slightly modified.)

Holliday junctions is a viable alternative if recombination is primarily or exclusively between sister chromatids.

b) Mathematical models of repeat tract expansion

Soon after the discovery of the nature of the DNA sequence responsible for fragile X syndrome, Morton & Macpherson (1992) proposed a model to explain in a Mendelian manner the pattern of genotypes and their inheritance found in a study of 33 families (Macpherson *et al.*, 1992). They defined four classes of allele: N (normal), S and Z, both defined initially as being in the range 150 - 400 bp (50 - 133 repeats) but which the family study suggested differed in their likelihood of conversion to L (full mutation alleles). They later said that from the results of Fu *et al.* (1991), S might be small premutations of 50-90 repeats and Z larger premutations of 90-200 repeats. They calculated mutation rates from each of these classes to the one above, taking 0 for the rate for Z $\rightarrow$ L in males and found a reasonable fit to observed frequencies *e.g.* of the penetrance of mental retardation in males or the probability that the sister of a Z male would have a full mutation and be mentally retarded, *etc.*. This did nothing to explain the mechanism of expansion but emphasized the observation that it tends to proceed in jumps rather than as a steady creep upwards. On their analysis of the fine structure of normal repeat alleles, Snow *et al.* (1994) proposed keeping this model with the addition of a first step of loss of AGG trinucleotides.

Di Rienzo *et al.* (1994) measured the frequencies of allele lengths of d(CA)·d(GT) repeats at ten different loci (overall range 74 - 201 repeats) in three human populations. They purposely chose loci which they hoped were not subject to selection which would bias the allele frequencies observed and concentrated on one of the populations for whom the demographic history had been investigated so that they could estimate frequencies with which alleles had common progenitors. They used computer modelling to compare the hypothesis of rare large jumps and more frequent single repeat changes with those of single-repeat changes only, or geometric

change, and found that the 'two-phase hypothesis' gave the better fit. Their model also seemed to fit the allele distribution of a 4-base repeat and they proposed that it would also apply to the relatively much less stable trinucleotide repeat sequences. It clearly agreed with the observations made on DM alleles (Imbert *et al.*, 1993) described earlier. Thus, though the disease-associated repeats might be more unstable, the expansion processes seemed likely to be similar.

c) Investigations of the mechanism of expansion

Strand *et al.* (1993) examined control of stability of $d[(GT)\cdot(AC)]_n$ tracts in centromere-containing plasmids in yeast and found that mutations in three mismatch-repair genes (*pms1, mlh1* and *msh2*) all caused 100 - 700-fold increases in instability of the tracts. One of their plasmids (with a tract of 33 bp in length) detected changes in repeat number that put an indicator, the *URA3* gene, out of frame and the other (with a tract of 29 bp) detected changes that put the gene into frame, so neither of them detected all the mutations in it. The authors found on sequencing mutations detected in wild type that 17/21 had only 1 or 2 repeats added (the remaining four having deletions of 5, 5, 7 and 8 repeats). When repeats were inserted into a chromosome rather than a plasmid the mutation rate was less but an increase of 5 repeats was seen in one case with no mismatch repair mutation. In the mismatch repair mutant strains *only* mutations of 1 or 2 repeats difference were found. The lack of more extensive changes in one of the strains (*pms1*), for which 36 mutants were sequenced, was statistically significant. The authors deduced that this indicated that the yeast *PMS1/MSH2/MLH1* system fails to recognize efficiently insertions or deletions that exceed 4 bp in size. They did not deduce that it might also suggest that an intact mismatch repair system is necessary for larger changes to take place. This has now been confirmed in this laboratory (Schmidt *et al.*, submitted).

Yeast has two polymerases that have 3'-5' exonuclease (proof-reading) activity (PolII and PolIII) and mutations in the nuclease portions of their genes had already been found to increase mutations in general by several hundred-fold but

Strand *et al.* (1993) found that these mutations had relatively very little effect on the repeat tracts, one increasing mutations 5-10-fold and the other not at all compared with wild type, so they deduced that either the mismatched bases were not at the end or the polymerases are insensitive to mismatches of more than one base pair.

Strand *et al.* (1993) then compared the rates of mutation in meiosis and mitosis since it had previously been found that recombination was about 100- to 1,000-fold more frequent in meiosis than in mitosis. They found that recombination in the heteroallelic *ura3* locus was increased 30-fold in meiosis without any increase in instability of the repeat tract. This confirmed their view that tract instability did not reflect 'normal recombination processes'. They pointed out that as reciprocal crossovers would produce an unstable dicentric plasmid their findings concerned only gene-conversion events. They did not consider the possibility that the gene-conversion events that they recorded elsewhere in the gene might have originated in the repeat tract and migrated.

They concluded that DNA polymerase *in vivo* has very high rates of slippage on templates of simple repeats but that most of these errors are corrected by mismatch-repair enzymes and that the instability of simple repeats in human diseases may be due either to increased DNA polymerase slippage or decreased mismatch repair. They pointed out that whether slippage results in increase or decrease in tract length may be affected by either the position of the unpaired bases (loop formed in the primer or template strand) or the frequency with which different types of errors (unpaired bases in the primer or template strand) are corrected but they did not consider how the human repeat expansions might be so much larger than the ones they had found.

As mentioned earlier, the suggestion that the mutational mechanism in fragile-X syndrome might be recombination (Pembrey *et al.*, 1984; Nussbaum *et al.*, 1986) came even before the locus was sequenced, as did the refutation of this that the mutation could occur in the absence of crossing over of flanking markers (Laird, 1987). Unequal crossing-over had several disadvantages as an explanation for

expansion of repeat sequences. Firstly, in unequal crossing-over one allele becomes longer but the other allele becomes shorter by the same amount, and no normal repeat allele had ever been observed to shrink. Secondly there were the findings of linkage disequilibrium. Kunst & Warren (1994) even went as far as to say that the linkage disequilibrium found between repeat alleles and marker loci by themselves for fragile-X and Imbert *et al.* (1993) for DM suggested that normal meiotic recombination was depressed in the vicinity of the repeat. Thirdly, the observation that expansion in the X-linked SBMA occurs more often in transmission by males (La Spada *et al.*, 1992) rules out meiotic recombination as a mechanism in that disorder. Fourthly, there was the observation that expansion was occurring in mitosis in some of the disorders, and finally, the maximum length of a new repeat tract generated by unequal crossing-over can only be a little less than the total length of the two parental alleles but sometimes the expansion was much more than this. However, recombination in other circumstances was not ruled out.

Jeffreys *et al.* (1988), in the study of the human 9 bp minisatellite repeat, MS1, mentioned earlier, looked at the frequencies of different sizes of repeat length change in paternally and maternally inherited alleles. They found that there was an excess of small changes, ≤ 10 repeats, in the paternal alleles and suggested that these were consistent with generation by replication slippage since ~400 cell divisions separate mature sperm from the male zygote whereas oocytes arise through only ~24 post-zygotic divisions. The many larger changes in repeat number, however did not show such a bias and the authors deduced that they were replication-independent and might arise by recombinational processes such as unequal exchange or gene conversion at meiosis.

Jeffreys *et al.* (1990) announced their idea that a good way of specifying the nature of an allele of a repeat tract that contained variant repeats was by the sequence not of its individual bases but of the interspersion pattern of variant repeats along its length. This could be done by the use either of restriction enzymes that would or would not cleave repeat variants (Jeffreys *et al.*, 1990) or of specific PCR primers

(Jeffreys *et al.*, 1991). Either method could be used to produce a gel pattern that could be read like a DNA sequence This was much less tedious than obtaining and comparing the base sequences and much more informative and accurate than merely comparing tract lengths.

The authors (Jeffreys *et al.*, 1990; 1991; 1994) studied the human locus of 29-bp-repeats, MS32, mentioned earlier. The repeats have two single-base substitution sites making 4 types of repeat. Jeffreys *et al.* (1994) looked for mutations in single sperm by PCR. An individual with 2 short alleles (of 42 and 63 repeats with different structures) was selected and typed for variant repeat structure and for a heterozygous 5' flanking marker. 90% of sperm which gained repeats showed gain within a few repeats of the 5' end. In a few cases these were simple reduplication of a segment of the progenitor allele consistent with unequal sister-chromatid exchange or replication slippage. However, in most cases the structure of the mutant allele was complex, containing sections from the other allele inserted into the progenitor, or both of the above and, in many cases, some unit or repeat blocks with no obvious origin, indicating further shuffling. Some deletion mutants were also complex with some repeats from elsewhere inserted into a deletion. In only one case out of 59 was the flanking marker exchanged.

Five mutations at the same locus in blood cells showed much simpler structures with no polarity and no segments of unknown origin. At the same locus, the authors investigated blood and sperm of another individual, with two flanking markers, with similar results, and looked for mutations in families and found the same structures in mutations of paternal origin but in mutations of maternal alleles only expansions that could be accounted for by simple duplication within one allele. Two other loci gave similar results.

Jeffreys *et al.* (1994) suggested that mutational polarity implied that mutation was modulated by element(s) outside the array and proposed activation of the recipient locus by introduction of a double-strand break by a protein binding to a mutation-initiator sequence 5' to the repeat array, with the position of the break

controlled by the initiator. 5' ends would then be taken back by exonuclease activity, leaving a gap with 3' overhangs. The gap would then be bridged by strand invasion from either the sister chromatid or the other allele followed by repair and resolution of the conversion complex. The recombination intermediates had to be resolved by disengaging the strands rather than by cleavage of Holliday junctions to account for the absence of crossing-over, as mentioned earlier. Explanation of the more complicated mutations required further breaks. The authors were less certain about how the deletions happened. They thought that conversion hot-spots might indicate that the mutation initiators were the promoters of synapsis.

Buard & Vergnaud (1994) carried out a similar study using their CEB1 human minisatellite but amplified the locus from diploid genomic DNA of individuals (presumably from blood) and separated the alleles by electrophoresis. They chose just three of the eight dimorphic positions within the 37 - 43 bp repeat unit to type the units and compared the variant repeat sequences of children with those of their parents. None of the mutations were associated with crossing-over. $^{2}/_{3}$ were expansions and of them $^{3}/_{4}$ were intra-allelic and $^{1}/_{4}$ interallelic rearrangements. The intra-allelic rearrangements were seen throughout the tracts but the interallelic ones were clustered near one end. In both cases complex rearrangements were seen, involving duplications and deletions within larger duplications amongst others. Of 18 contractions investigated, all were simple deletions except one that involved loss of 15 repeat units and gain of two.

The authors suggested that all the expansion rearrangements could be explained by one of the parental alleles being broken not by a blunt-ended double-strand break but by nicks in each strand several repeat units apart, followed by separation of the two parts of the allele, each with a long 3' overhang, and recombination with the other allele or sister chromatid. They observed that it was very striking that in all of the interallelic exchanges the beginnings of the two parental alleles were exactly aligned and that this also applied to the interallelic exchanges reported by Jeffreys *et al.* (1994). They pointed out that this was achievable most

easily with the first few motifs and that the overall size difference usually observed between alleles made complete pairing impossible but with sister-chromatid exchanges alignment could be made all along the allele. This explanation of clustering of interallelic exchanges did not require a *cis*-acting element. Deletions, they suggested, might result from the initial nicks giving 5' overhangs which were then trimmed back to 3' ones before recombination, or sometimes from single-step intramolecular events.

Jansen *et al.* (1994) made a study of repeat lengths at the DM locus in myotonic dystrophy of varying degrees of severity. They compared repeat lengths between sperm and a wide variety of somatic tissues within individuals and between twins and between fathers and offspring. They concluded that, in addition to initial repeat-tract length, the number of cell divisions involved in tissue formation, and perhaps a specific selection process in spermatogenesis, might influence the dynamics of expansion, and proposed that the mechanism of trinucleotide repeat expansion is similar to repair of deletions and double-stranded gaps.

They proposed that unequal pairing of sister chromatids occurred with the formation of a four-stranded synaptic structure and that in repeats over a critical length staggered cruciforms, triple-stranded structures or loops in the double-stranded helices of either chromatid might form. Double strand cleavage by endonuclease of strands *opposite* to the looped structure was then envisaged with subsequent repair of the gap, using donor strands as a template, leading to expansion, *i.e.* a sort of double-stranded version of the mechanism proposed by Levinson & Gutman (1987) and Petruska *et al.* (1998). If this occurred during consecutive mitotic divisions, they suggested, new alleles of more than the sum of the parental allele lengths could appear. Alternatively, if the secondary structure was cleaved and removed deletion would result.

Jansen *et al.* (1994) also suggested that 'VSSMs' (very simple sequence motifs), depending on their length, might be arrest sites for replication. If elaborate mispaired structures had to be resolved before mitosis could proceed this scheme

could explain the differential effects on repeat tracts in male and female transmission because spermatogenesis requires more rapid cell divisions than does oogenesis. They mentioned that others had proposed that such effects could underlie the restriction of congenital myotonic dystrophy cases to maternal transmission, but neglected to mention that in other conditions juvenile onset cases come only from *paternal* transmission. They also revealed that in a case of contraction of a paternal disease allele into the normal range at 24 repeat units, the recipient of that allele had passed it on stably to his first child, and commented that this argued against the existence of a neighbouring mutation-initiator element.

7. Concluding remarks

The genomes of multicellular eukaryotes abound with polymorphic tracts of repeated DNA sequences. Some of these show a balance of expansion and contraction mutations while others show a marked tendency to expand rather than contract, at least up until some possible upper limit at which contraction might become more common. Their mutation rates had been found to vary but small changes in length of one or two repeat units, explainable by replication slippage, and larger changes requiring some other explanation appeared to be common features and trinucleotide repeat tracts just seemed to be a part of the spectrum. The trinucleotide repeats that had so far been associated with inherited disorders were seen to have very high mutation rates when they exceeded some threshold length but were not the only repeat sequences to have high mutation rates. They were not even the only *tri*nucleotide repeats to expand and their notable association with disease seemed to be mainly because of their more frequent occurrence in genes than other repeating sequences. Several sequences with much longer repeating units had already been found to be associated with disease by 1994 and these will be mentioned in a later section.

The mechanisms suggested for trinucleotide repeat expansion discussed above had all been previously suggested or demonstrated for other types of repeating

sequences. It could be however that all of the sequences with very high mutation rates might be so unstable because one or both strands had a tendency to form secondary structure. Secondary structure might aid replication slippage by larger numbers of repeat units more often and it could also engender more frequent recombination by providing sites for DNA nicks or breaks.

One difference however in expansion of different repeat tracts was that expansion of trinucleotide repeats had been found to be dependent upon the number perfect repeats whereas the minisatellites studied by Jeffreys *et al.* (1994) and Buard & Vergnaud (1994) were highly unstable despite sequence differences between repeat units. This presumably was because repeat units of around 30 bp in length still had enough similarity to pair despite some differences whereas with 3 bp units a single base-pair difference is a major difference. The reports of interrupted fragile-X repeat tracts (Kunst & Warren, 1994; Snow *et al.*, 1994; Hirst *et al.*, 1994) showed that there could be one, two, three, or occasionally four roughly evenly-spaced d(AGG)·d(CCT) interruptions, depending upon the length of the normal allele, indicating that sometimes there must have been an expansion by a 30 bp unit, d[AGG$(CGG)_9$·$(CCG)_9$(CCT)], or circular permutation thereof. As no more than four such ~30 bp units have been recorded in a tract, these changes would not be expected to occur very often. However, a tract with many such 30 bp units might be just as unstable on this scale as the pure trinucleotide repeats are on their scale.

Finally, it was clear that major repeat instability is normally restricted to meiosis and the first few mitotic divisions thereafter, and may be much more in one sex than the other, and therefore other factors must be involved. Eichler *et al.* (1994) suggested that this might either be depletion of ATP pools (known to cause shortening in average Okazaki fragment lengths) or delay in activation of embryonic methylases. That other factors must be involved was also brought out by the fact that the same sequences that tend to expand in multicellular eukaryotes tend to contract when cloned in bacteria.

One expansion mechanism that has not been mentioned so far was proposed by Richards & Sutherland (1994). They estimated that the large expansions seen in DM and fragile-X syndrome started to occur when the repeat copy number was greater than about 80 and suggested that this might correspond to the repeat tract exceeding the length of an Okazaki fragment. [The length of an Okazaki fragment is 25 - 300 nt in mammals (DePamphilis & Wassarman, 1980)]. Richards & Sutherland (1994) proposed that if the repeat tract was shorter than an Okazaki fragment the 5' end of the fragment would be 'anchored' by unique sequence and only simple slippage would occur, but that as the repeat tract grew in length there would be an increasing chance that two single-strand breaks (the ends of an Okazaki fragment) would occur within the repeat tract. A fragment that was composed exclusively of trinucleotide repeats might, they proposed, be able to 'slide' on its template - the 3' end of the fragment coming to rest on a more 3' part of the template and the 5' end of the fragment melting off and waving freely about the upstream double-stranded unique DNA - and that subsequent repair would lead to expansion.

Neither the mechanism of sliding, nor the mechanism by which repair would lead to expansion was discussed. Concerted melting of a whole Okazaki fragment and its reannealing in a new position seemed implausible. Eichler *et al.* (1994) suggested that expansion might occur by intra-strand pairing of the Okazaki fragment with movement not only of its 3' end but also its 5' end towards the middle of the repeating template. We suggested (Darlow & Leach, 1995) that another possibility might be that the threshold length for instability might be related not to the length of an Okazaki fragment but to the minimal length of homology required to initiate homologous recombination. This minimal efficient processing segment (MEPS) (Shen & Huang, 1986) is between 200 and 300 base-pairs in mammalian cells (Rubnitz & Subramani, 1984; Ayares *et al.*, 1986; Liskay *et al.*, 1987). For a repeat tract below this length, the pairing of the broken chromosome would have to rely on homology outside the repeat array. This would anchor the event and prevent significant changes in number of repeats. On the other hand, an array of repeats

longer than the MEPS could recombine without external anchoring and lead to more frequent and variable changes in numbers. The observation that partial sequence divergence severely inhibits recombination in mammalian cells (Waldman & Liskay, 1987) would also account for the suppression of instability by imperfect repeats (Chung *et al.*, 1993; Kunst & Warren, 1994; Snow *et al.*, 1994; Hirst *et al.*, 1994). This mechanism would not require the formation of large secondary structures because even a small structure could be cleaved and lead to recombination.

The formation of DNA secondary structure had thus been implicated in most proposals for the mechanism of expansion of trinucleotide repeat arrays. The aim of this research project was to show whether trinucleotide repeat tracts might form secondary structure *in vivo*. The basis of the work is described in the last section of this chapter after a review of some of the developments in the literature during the course of the work.

Further developments

The range of research

There are now several thousand papers on repeat expansion disorders and it is not intended to give a detailed cover of the field here. There are already many reviews but even these tend nowadays to be confined to particular topics.

Publications have included: more searches for trinucleotide repeats in genomes; reports of more repeat expansion disorders and loci (see below); many unsuccessful attempts to find trinucleotide repeat expansions in other disorders, particularly psychiatric conditions; work on DNA secondary structure *in vitro* and on expansion *in vitro* and *in vivo*, in *E. coli*, yeast, and human cell lines, along with revised theories on the mechanism(s) of expansion (to be discussed in later chapters); work and theories on the mechanisms by which the expanded tracts lead to pathology (reviewed in Klockgether & Dichgans, 1997; Zoghbi, 1997; Koshy & Zoghbi, 1997; Kakizuka, 1997; Koeppen, 1998); relationship of clinical features to repeat tract

length and the possible existence of modifying genes; papers on population genetics and evolution of trinucleotide repeat tracts, some dealing with tract lengths in different human populations and other species, some with repeat interruptions, and some with segregation distortion (the preferential transmission of the longer, or in some cases the shorter allele); analyses of instability of repeats through studies of individual sperm and ova, including effects of age of the parent, possible *cis*- and *trans*-acting factors affecting likelihood of expansion, and continuing debate over whether expansion is prezygotic or postzygotic; studies of protein binding by trinucleotide repeats in DNA and RNA, and methylation; mouse models of trinucleotide and other repeat expansion (reviewed in Korneluk & Narang, 1997; Longo & Massa, 1997; Burright *et al.*, 1997) and recently a transgenic *Drosophila* with part of the human *MJD1* polyglutamine repeat gene (see below) exhibiting a late-onset neurological disorder (Warrick *et al.*, 1998).

Unfortunately some of these topics are not individually reviewed but some recent general reviews that give some cover to several topics are Longshore & Tarleton (1996, more comprehensive than the others, though not so recent), Tsuji (1997), La Spada (1997), Reddy & Housman (1997), and Harper (1997).

More folate-sensitive fragile sites

Since the start of this work, three further folate-sensitive fragile sites have been cloned and all have d(CGG)·d(CCG) repeats. The first was *FRA16A* (Nancarrow *et al.*, 1994). Like *FRAXA* and *FRAXE*, this site was also found to be adjacent to a CpG island which is methylated in individuals expressing the fragile site but unlike them and in common with most other fragile sites, there was no phenotype associated with its expression. There is no methylation (imprinting) in this region when the fragile site is not expressed, suggesting that methylation is a consequence not a cause of expansion. The amplification observed in *FRA16A*, ~3 to ~5.7 kb is larger than for the X chromosome fragile sites perhaps because of lack of selection against large alleles. Like *FRAXA*, expansion was found to be associated with loss of

an interruption (Nancarrow *et al.*, 1995). The sequence on the C-rich strand is $d(CCG)_{3 \text{ or } 6}\ (CCT)_{1 \text{ or } 2}\ (CCG)_{6-9}\ (CTG)_{0 \text{ or } 1}\ (CCG)_{6-11}\ (CCTCCA)_{1 \text{ or } 2}$. except that some normal alleles without the CTG have more than 20 d(CCG) repeats before the terminal section of the tract. All of the expanded alleles lack the CTG interruption. The locus was found to be highly polymorphic only in European populations (Nancarrow *et al.*, 1995; Richards & Sutherland, 1997). Richards *et al.* (1997), who analysed the distributions of copy number polymorphisms at 12 trinucleotide repeat loci, concluded that this indicated that a *cis* component is important in the mutation mechanism at *FRA16A* rather than *trans*-acting factors that would affect other loci.

FRAXF was discovered after it was found that in some families with fragile sites at Xq27.3-q28 did not have expanded repeats at *FRAXA* or *FRAXE*. *FRAXF* was reported to be $d[(GCCGTC)_m(GCC)_n(GGC)_n(GACGGC)_m]$ with m = 3 or 4 (Parrish *et al.*, 1994; Ritchie *et al.*, 1994) and a total length in individuals not expressing the fragile site of 6 - 38 triplets (Holden *et al.*, 1996) and expanded alleles in the range 300 - >900 repeats (Parrish *et al.*, 1994; Ritchie *et al.*, 1994). Expansion and methylation of the *FRAXF* site was demonstrated in mentally retarded and mentally normal individuals within the same families (Parrish *et al.*, 1994; Ritchie *et al.*, 1994) but only a few families with the expansion were found and it was not possible to say whether this reflected a genuine association between the site and a gene related to mental retardation with mosaicism causing variable penetrance, or whether there was only an apparent association due to ascertainment bias. However, the repeat tract was found to be in an open reading frame in both strands and one of these would encode a protein with significant homologies to the mouse *engrailed-1* homeobox gene which has a polyalanine tract encoded by the repeat. Surprisingly, nothing further has yet come to light on these possible genes in humans. The human *EN1* and *EN2* genes are on chromosomes 2 and 7 respectively (Logan *et al.*, 1989).

There appears to be some connection between expansion at these three fragile sites on the X chromosome. Brown *et al.* (1997) found no correlation between normal *FRAXA* and *FRAXE* repeat sizes but, though they found no cases of *FRAXE*

expansion in 953 individuals examined, one case of *FMR2* deletion, two cases of *FRAXE* instability and one *FRAXE* mosaic male were all found amongst individuals who had expansion of *FRAXA*. Furthermore, significantly larger *FMR2* alleles segregated with expanded *FMR1* alleles. Added to this, Barnicoat *et al.* (1997) found that in three *FRAXE* families, all the affected males also had large methylated repeats at *FRAXF*.

Jacobsen syndrome is due to deletion of the end of the long arm of chromosome 11 (11q23→qter). Jones *et al.* (1994) reported a family in which the normal mother and brother of an affected individual expressed the rare folate-sensitive fragile site, *FRA11B*. The breakpoint of the deletion in the affected boy was mapped to the region of the fragile site, one of the genes mapped to the area, the proto-oncogene *CLB2*, was known to have a d(CCG)·d(CGG) repeat in the 5'-untranslated region, and four of the clones of the fragile site region were found to have expanded d(CCG)·d(CGG) repeats. Jones *et al.* (1995) then showed all individuals expressing the fragile site in six families had the expanded repeat, with >100 copies, in *CBL2* and a nearby CpG island was found to be methylated but presumed loss of function of one of their *CBL2* genes did not appear to have any phenotypic effect.

Four further unrelated cases of Jacobsen syndrome were examined and one of them had the expansion (his mother having a premutation of 85 repeats and not expressing the site) but in neither this case nor the original one was the *CBL2* gene disrupted. Probes indicated that the breakpoint was within about 20 kb proximal to the fragile site. The authors (Jones *et al.*, 1995) pointed out that both *FRA11B* and Jacobsen syndrome are rare and that the association was unlikely to be chance though the chromosome break did not occur at the fragile site.

They subsequently reported two other cases of Jacobsen syndrome (Michaelis *et al.*, 1998) in which there was no evidence of expansion of the *CBL2* repeat and the deletion breakpoint was approximately 1·5 - 3 Mb telomeric to *FRA11B*. They suggested that these findings and those in the previously reported patients indicated that the breakpoint for most 11q deletions in Jacobsen syndrome is

telomeric to *FRA11B* and raised the possibility that there may be other fragile sites in 11q23.3 in addition to *FRA11B*. Horwitz *et al.* (1996) consider *CBL2* as one of three possible candidate genes to account for anticipation in familial leukaemia but this seems a little unlikely since *CBL2* acts as an oncogene in a dominant manner by overexpression and expansion of the d(CCG)·d(CGG) repeat is expected to cause underexpression.

FRA11B brings the current number of folate sensitive fragile sites shown to be expanded d(CCG)·d(CGG) repeat tracts to five but a recently reported d(CCG)·d(CGG) tract in the same region of the X chromosome as *FRAXA*, *E* and *F* but not so far seen expanded (Ritchie *et al.*, 1997) may turn out to be another.

More CAG/polyglutamine repeat disorders

To SBMA, HD, SCA1 and DRPLA four more degenerative neurological disorders caused by expanded d(CAG)·d(CTG) repeat tracts coding for polyglutamine have been added. Haw River Syndrome proved to be due to the same expansion as DRPLA but in a different genetic background (American Negroes, Burke *et al.*, 1994) but new ones were Machado-Joseph disease (MJD, Kawaguchi *et al.*, 1994), now classified as Spinocerebellar Ataxia Type 3 (SCA3), followed by SCA2 (Pulst *et al.*, 1996; Sanpei *et al.*, 1996; Imbert *et al.*, 1996), SCA6 (Zhuchenko *et al.*, 1997; Riess *et al.*, 1997), and SCA7 (David *et al.*, 1997).

The normal SCA2 tract has 1 - 3 d(CAA)·d(TTG) interruptions with 4 - 5 d(CAG)·d(CTG) repeats between them whereas in patients the tract was found to be perfect. The normal range reported is 14 - 29 units, with about 72% of individuals having 22 and most of the rest having 23, and the disease range reported is 34 - 64 (Pulst *et al.*, 1996; Sanpei *et al.*, 1996; Imbert *et al.*, 1996; Lorenzetti *et al.*, 1997; Cancel *et al.*, 1997), though Leggo *et al.* (1997) have reported an individual with 33 repeats who is so far unaffected.

In SCA3 the sequence at the beginning of the repeat tract in the coding strand is CAG CAG CAA AAG CAG CAA but after that the tract is pure CAG repeats

and the variants are not lost in the expansion, which occurs in the pure part of the tract (Kawaguchi *et al.*, 1994). The normal range was reported as 13 - 36 and disease range 68 -79 (Kawaguchi *et al.*, 1994) but males have an age of onset about 5 years earlier than females with the same length of tract (Kawakami *et al.*, 1995).

In SCA6 there are no interruptions and much smaller numbers of repeat units are involved. The normal alleles have 4 - 20 units and the disease range was found to be only 21 - 30 with a sharp negative correlation between repeat number and age of onset (Zhuchenko *et al.*, 1997; Ishikawa *et al.*, 1997; Matsuyama *et al.*, 1997; Riess *et al.*, 1997; Ikeuchi *et al.*, 1997). The alleles in the disease range were found to be fairly stably transmitted. Matsuyama *et al.* (1997) noted an increase from 24 to 26 units in one out of eight parent-child pairs examined and Ikeuchi *et al.* (1997) found two siblings differing by one repeat unit (22 and 23) in one of the 30 families they examined. Ishikawa *et al.* (1997) found no change in 15 pedigrees with numerous affected individuals and Riess *et al.* (1997) found no change in 11 parent-child pairs. This condition illustrates that the number of glutamine residues required to cause disease depends upon the particular gene and is not related to the number of trinucleotide repeats required to cause instability. Some of these investigators (Matsuyama *et al.*, 1997; Riess *et al.*, 1997; Ikeuchi *et al.*, 1997) found anticipation despite the fact that the repeat size did not change between generations but none of them appeared to realize this would be due to precisely the kinds of statistical bias that Penrose had pointed out are bound to occur in data on disorders with variable age of onset.

In SCA7 the normal range reported is from 4 (Del-Favero *et al.*, 1998) to 35 (David *et al.*, 1998) and the disease range from 34 (Gouw *et al.*, 1998) to >200, with the longest polyglutamine tract ever reported (Johansson *et al.*, 1998). (The latter patient died at 7 months old.) The instability has been found to be greater on paternal transmission, with increases of 15 ± 20, than in maternal transmissions with increases of 5 ± 5 repeat units (David *et al.*, 1998) yet despite this, maternal transmission of the disease is more common (Gouw *et al.*, 1998). Somatic mosaicism

in leukocyte DNA suggests that expanded SCA7 alleles are unstable in mitosis (Gouw *et al.*, 1998).

In DRPLA (Sato *et al.*, 1995), SCA3 (Kawakami *et al.*, 1995; Takiyama *et al.*, 1995; Sobue *et al.*, 1996) and SCA6 (Matsuyama *et al.*, 1997; Ikeuchi *et al.*, 1997; Ishikawa *et al.*, 1997) homozygotes for expanded alleles tend to have earlier ages of onset than heterozygotes with the same repeat numbers. In HD (Snell *et al.*, 1993) and SCA3 (Dürr *et al.*, 1996) even the number of repeats on the normal allele has been shown to influence the onset age, and in SCA3 two homozygous individuals were reported with a mild form of the disease and a tract length (41 or 40 units) that does not cause disease in the heterozygous state, *i.e.* an intermediate length of tract was found to be recessive (Kurohara *et al.*, 1997).

Non-pathogenic expanding d(CAG)·d(CTG) repeats

Given the large number of searches for trinucleotide repeat expansions in diseases of variable age of onset which appeared to show anticipation and the finding of expansions of d(CAG)·d(CTG) repeats amongst others in a general population screen by Lindblad *et al.* (1994) by their RED method, it was inevitable that some searches would locate expanded repeats that were not obviously associated with disease.

Because of reports of linkage of bipolar affective disorder to chromosome 18, Breschel *et al.* (1997) screened a chromosome 18 library and found a clone with $d[(CAG)\cdot(CTG)]_{24}$ in an intron of the *SEF2-1* gene which encodes a DNA-binding protein involved in transcriptional regulation. CTG is on the coding strand. Screening of bipolar disorder families and CEPH reference families revealed that the repeat is highly polymorphic with relatively stable alleles in the range 10 - 37 trinucleotides, moderately enlarged and unstable alleles of 53 - 250 repeats - 3% of both groups - and very enlarged alleles, with 800 - 2,100 repeats, in three unaffected healthy individuals in the bipolar disorder families. Two individuals were found to be mosaics having three alleles all in the normal range.

Nakamoto *et al.* (1997) searched for expanded d(CAG)·d(CTG) repeats by RED in patients with familial spastic paraplegia and Holmes-type cerebellar ataxia. From patients with each disorder they isolated clones with pure expanded repeat tracts and were surprised to find that they were from the same locus on chromosome 17. A screen of 75 members of the Japanese population and 30 Caucasians revealed two clusters of allele sizes, 10 - 30 and 55 - 92 trinucleotides. The larger alleles were found to be stably transmitted, though within the range of unstable alleles in trinucleotide repeat disorders. The flanking sequences showed no homology with any reported sequences and the authors suggested that this might indicate that transcription is necessary for instability. There is some evidence from work in *E. coli* that transcription may increase instability (Bowater *et al.*, 1997) though clearly it is not essential unless all the intergenic expanding repeats are in fact transcribed.

A d(GAA)·d(TTC) repeat disease

Friedreich's ataxia (FRDA), the most common hereditary ataxia, with a prevalence of about 1 in 50,000, was shown to be due to mutation of a gene, *X25*, coding for a protein which was named frataxin (Campuzano *et al.*, 1996). The inheritance is autosomal recessive. In a few alleles there were point mutations but in about 98% the mutation was expansion of a d(GAA)·d(TTC) repeat tract in an intron. The tract was found to be polymorphic in normal alleles with a range of 7 - 22 trinucleotides, and expanded alleles carried 200 - >900 repeats and were unstably transmitted. No instances of expansion of a normal allele into the disease range or of reversion of an expanded tract into the normal range were seen and the authors pointed out that as the disease was recessive, expanded alleles could be maintained in the population without being subject to selection in the heterozygous state. Subsequently the normal range has been extended up to 34 repeat units, the majority of large (16-34-unit) normal alleles have been found to have the same haplotype as the majority of expanded alleles, and two premutation alleles of 42 and 60 units have

been recorded that expanded into the disease range in a single generation (Cossée *et al.*, 1997).

Campuzano *et al.* (1996) pointed out that d(GAA)·d(TTC) repeats, up to 30 - 40 units, are common in many organisms and are sometimes polymorphic, as in the 3'-untranslated region of the rat polymeric immunoglobulin protein. (This presumably reflects the great increase in the size of sequence databases since the time of Stallings' (1994) paper.) McMurray and colleagues, (Gacy *et al.*, 1995) had suggested that hairpin formation was the explanation for d(CXG) repeats being responsible for all trinucleotide repeat disorders to that date and had pronounced that d(GAA)·d(TTC) repeats formed no secondary structure and implied that they would therefore not be likely to expand. Campuzano *et al.* (1996) said that on this basis their own finding was unexpected.

In an editorial introducing the paper of Campuzano *et al.* (1996), Warren (1996) seemed to think that their finding radically altered the picture of repeat expansion disorders. He said that d(CXG) repeats no longer appeared unique, apparently completely unaware of the finding of Lindblad *et al.* (1994) that six trinucleotide repeats expand in humans including d(GAA)·d(TTC). He said that dominant inheritance was no longer a valid characteristic of expansion disorders, ignoring the fact that SBMA, one of the first two diseases found to be due to trinucleotide repeat expansion (La Spada *et al.*, 1991), is recessive (X-linked but nevertheless recessive, unlike fragile X which is X-linked dominant). He even went as far as to call into doubt the importance of secondary structure in the mechanism of DNA repeat expansion.

We, however, were unperturbed by the implication that the purpose of this project might be rendered meaningless. Not only had Lindblad *et al.* (1994) shown that d(GAA)·d(TTC) repeats could be found in expanded lengths in some people but Epplen *et al.* (1991) had already found that these tracts were highly unstable in hens. Lee (1980; 1990) had shown that long poly-d(GAA) strands form secondary structures, thought to be tetraplexes, and Epplen *et al.* (1991) had pointed out that

d(GAA)·d(TTC) repeats would be expected to form triplexes because of their polypurine·polypyrimidine nature. In fact, Wells and colleagues (Shimizu *et al.*, 1989; Hanvey *et al.*, 1989) had shown this to be the case. (However, McMurray, Morton-Bradbury, Gupta and colleagues (Gacy *et al.*, 1998; Mariappan *et al.*, 1999) have now published this almost as though it were a new discovery.)

Coding d(CGG)·d(CCG) repeat expansion diseases

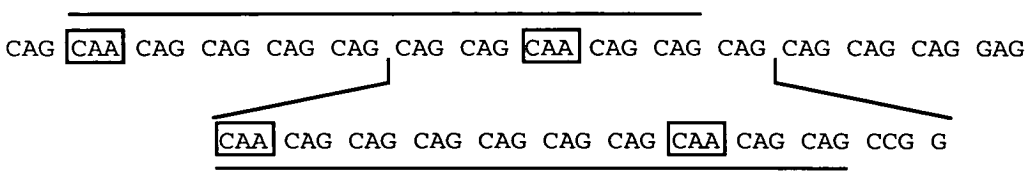
Reddy & Housman (1997) divided the trinucleotide repeat disorders into 'Type I', in which the repeats are coding and are all CAG repeats which expand to a limited extent, and 'Type II' which are non-coding, more varied - CGG, CTG, GAA - and expand much more. Since Stallings (1994) had revealed that a large proportion of d(CGG)·d(CCG) repeat tracts are in exons and Green & Wang (1994) had shown that these are at least occasionally coding, this classification could have been expected not to last; it was already out of date before it was published. Just as d(CAG)·d(CTG) tracts may or may not (in the case of DM) be coding, so d(CGG)·d(CCG) repeats causing disease may not or may be coding.

Muragaki *et al.* (1996) first reported that synpolydactyly (SPD; extra and fused fingers and toes) can be due to expansion of a tract in the *HOXD13* gene coding for polyalanine. Most of the codons are GCG but some are GCA, GCT, or GCC, so it can be observed exactly which section of the tract has been duplicated. The normal protein contains 15 alanine residues and was not noted to be polymorphic. In three families the expansion was a duplication of a 7-, 8-, or 10-trinucleotide section of the tract, starting at different points. In one of these the expansion was a new mutation and in the other two the expanded allele was stably inherited for, in one family, at least six generations. The expression is dominant with incomplete penetrance, two cases of carriers being found without the phenotype. Subsequently, with others (Goodman *et al.*, 1997), the authors reviewed these families along with 17 more families with tracts of up to 29 alanine residues. They found a striking increase in penetrance and severity of phenotype with increasing expansion but again all of the

expansions were single duplications of sections within the tract and all were stably inherited. Note that this disorder is the first trinucleotide repeat disorder not to be a neurological degeneration but a developmental malformation.

Cleidocranial dysplasia (CCD) is another dominantly inherited developmental malformation (of the skull and collar bones but also involving short stature and other changes in skeletal patterning and growth). Mundlos *et al.* (1997) have shown that is due to mutations in *CBFA1*, a member of the runt family of transcription factor genes. In some families the phenotype segregates with deletions resulting in heterozygous loss of the gene. In other families, insertion, deletion or mis-sense mutation lead to stop codons, but in one family in-frame expansion of a tract coding for polyalanine results in minor clinical findings of CCD including short fingers. Most of the codons are GCG but there are two GCT interruptions (separated by 5 GCG codons) and the tract ends GCA GCT GCA. In most normal individuals the tract has 17 codons but in 7/320 alleles (160 normal individuals) there was an allele with a 6-codon deletion, GCG GCG GCG GCG GCG GCT or a circular permutation thereof. In the particular family concerned, with 27 members including 16 affected of whom 6 were available for testing, the normal 17-residue tract was increased by a ten-codon expansion which is a direct repeat of a central part of the tract, $(GCG)_2$ GCT $(GCG)_5$ GCT GCG (or one of some of the circular permutations thereof). Thus expansion has occurred despite the interruptions.

Another mutation, one of the frame-shift mutations in the same gene in one of the other families, is also relevant because it involves mutation of a polyglutamine-coding tract. The mutation is reproduced below as the authors have represented it:



In this portrayal, one section of the tract is replaced by a longer one which appears to be a copy of an overlapping section, indicated by the long line, plus the replacement

of a CAG codon by the tetranucleotide, CCGG. An alternative portrayal would be of direct duplication of the first five codons lined, CAA $(CAG)_4$, [or, indeed, of the section CAG CAA $(CAG)_3$], and a separate replacement of A by CG further down the tract. Either way, this event seems to cry out for an explanation involving recombination, not slippage. It was seen as a new mutation in a single affected individual.

Another disorder due to expansion of a polyalanine tract (in the poly(A)-binding protein 2 gene, *PABA2*) is occulopharangeal muscular dystrophy (OPMD, Brais *et al.*, 1998). The normal allele encodes a run of 10 alanine residues of which the first 6 are encoded by $(GCG)_6$ and all expansions were seen in this part of the DNA tract. An allele with $(GCG)_7$ was found in 2% of French Canadian control chromosomes and in affected families was found to cause the disorder recessively. Longer alleles, with $(GCG)_{8-13}$ caused the disorder dominantly and four compound heterozygotes, all $(GCG)_7/(GCG)_9$, had a more severe phenotype than $(GCG)_6/(GCG)_9$ individuals. (The codon preceding the GCG codons is ATG and the codon following them is GCA, so the DNA repeat tract $d[(GCG)\cdot(CGC)]_6$ is actually part of a longer repeat, $d[(GGC)\cdot(GCC)]_7$, and could also be described as $d[(CGG)\cdot(CCG)]_6$.) As in the other disorders, the expanded alleles were inherited fairly stably. A single mutation, from, in codon terms, $(GCG)_9$ to $(GCG)_{12}$, was observed in what the authors estimated from their pedigrees to be about 600 meioses.

In all of the examples of expansions in tracts coding for polyalanine the expanded alleles are stably inherited, but so are the disease-causing alleles in SCA6. This begs the question of exactly what we term as a repeat expansion disorder. (The reason for not using the expression 'trinucleotide repeat disorder' is that apart from the one case of dinucleotide expansion causing an oncogenic mutation mentioned earlier (Fearon *et al.*, 1990), expansion of tracts of repeat units of other lengths has also been observed to cause disease.) Just as repeat sequences in intergenic regions vary widely in their stability, so do repeat sequences within genes, but the latter can

cause inherited disorder when they mutate whether mutation is frequent or infrequent.

Other disease- or fragile-site-causing repeats

Other types of fragile site

Apart from the folate-sensitive group, there are two other types of rare fragile site, distamycin-A-inducible and bromodeoxyuridine-requiring (Sutherland & Richards, 1995). *FRA16B* became the first non-folate-sensitive rare fragile site to be sequenced (Yu *et al.*, 1997a). (The one with d(CGG)·d(CCG) repeats was *FRA16A*.) It belongs to the group of sites that are seen when the cells are grown in the presence of distamycin A, or other compounds such as berenil and netropsin, that are known to bind to AT-rich sequences (but also with bromodeoxyuridine). It might have been expected therefore that it would be an AT-rich repeating sequence, and so it proved. The locus in people not expressing the fragile site was found to consist of 7 - 12 copies of an imperfectly repeated sequence of 26 - 33 bp units giving at least 13 different alleles. Comparison of the units showed both terminal and interstitial deletions in the shorter units but all near one end of the unit. λ clones of even these non-fragile-site-producing alleles were unstable in *E. coli*.

The fragile site was found to consist of 33 bp repeats of which the sequence on one strand is d(ATATATTATATATTATATCTAATAATATAT$^{C}/_{A}$TA). Both versions of this repeat occur in the normal sequences. The whole region consists almost entirely of AT-rich DNA with several types of repeated sequence including two other short minisatellites with 32 and 37 bp repeat units respectively. The identity of the particular repeat as the fragile site was confirmed by PCR using primer pairs consisting of either strand of one copy of the repeat and a unique proximal or distal flanking primer. Ladders of evenly-spaced bands were then seen on electrophoresis. The fragile site consisted of up to 2,000 repeat units. It has subsequently been shown (Hewett *et al.*, 1998) that though this 'ladder PCR' is

effective in demonstrating the presence of expanded repeats, it is not specific enough to identify the exact sequence of the repeat motif, so there may be more variation in the repeating motif than just a single base substitution, but all ladders have given equal spacing of bands.

Instability of *FRA16B* within families was not detected. The authors accounted for this by explaining that the smallest expanded repeat was ~15 kb and the restriction-fragment containing it >20 kb, running within the zone of compaction on ordinary agarose gels. On pulsed-field gel electrophoresis no instability was detected either but the authors said that even several kilobases would not have been resolved under these conditions. However, since rare fragile sites are generally inherited in a Mendelian manner, intermediate ('premutation') alleles are likely to be much rarer than the fragile sites themselves and transition between normal and expanded alleles in either direction very uncommon. Indeed, since the same repeat was found to be responsible for *FRA16B* in every family examined, it may be that all the expanded alleles derive from a single distant expansion event from the normal range. Also, size changes may be less frequent than with the folate-sensitive fragile sites because of secondary structures in AT-rich DNA being less stable than ones in GC-rich DNA.

Though differences in size within families were not detected, differences between families were always seen. One individual in a *FRA16B* family was found to have a faint ladder of bands but was below the threshold of cells expressing the fragile site (2%) indicating that he is a mosaic for the fragile site and therefore that there can be somatic instability sufficient to reduce an expanded repeat down to a length that does not express. Yu *et al.* (1997a) cited one case from 1983 of a family in which the father expressed *FRA16B* and a child had a chromosome 1;16 translocation with the break-point on chromosome 16 mapping to the region of *FRA16B*, indicating the possibility of breakage related to the fragile site as at *FRA11B*.

More recently a rare fragile site of the bromodeoxyuridine-requiring group has been sequenced (Hewett *et al.*, 1998). It is only induced by bromodeoxyuridine or

bromodeoxycytidine, not by distamycin A *etc.*, but was also found to be very AT-rich (91% in the central repeat units). The repeat units vary in length between 16 and 52 bp and differ between the peripheries and the interior of the tract. The investigators divided the alleles into four classes: short- intermediate- and long-normal, and *FRA10B*-expressing and, from analysis of their sequence compositions, suggested that they had arisen, each from the previous class, in this order. Short normal alleles (~66% of the total) had a particular set of proximal repeat units and only four or five of the internal units seen in expanded alleles. The intermediate normal alleles (~33%) also had certain unique flanking repeat units and the long-normal (<1%) and expanded alleles shared a subset of flanking units (one of which was also seen in short alleles). Expanded alleles from different families had different variant repeats in the expanded portion but were the same within families. Assuming an average length of 42 bp, the threshold for instability appeared to be about 75 repeat units. I shall return to this in the final chapter. The authors note that like *FRA16B* repeats, the *FRA10B* repeats could form hairpins. They also suggest that bromodeoxyuridine may have a specific sequence requirement. If so, since this compound can also induce Distamycin-A-inducible sites, it would seem to be a less specific agent than Distamycin A. Distamycin A binds to AT-rich runs of ≥3 bp provided that there are no TpA steps, and it does not bind to normal B-DNA but only to the narrow groove of B'-DNA (Yu *et al.*, 1997a).

Since common fragile sites, by definition, are present in all individuals, it is perhaps not too surprising that it has turned out that they are not expanded repeat sequences. Mishmar *et al.* (1998) have characterized and sequenced the aphidicolin-inducible common fragile site, *FRA7H*. It was found to have several regions with potential for unusual DNA structure, including high-flexibility, low-stability, and non-B-DNA-forming sequences. The region was found to be 58% AT-rich and contains 13·1% short interspersed elements, 13·8% long interspersed elements, 5% long terminal repeats, and 0·7% DNA transposons, but no expanded repeats were

found. The partial sequence of *FRA3B* and putative partial sequence of *FRA7G* were found to be similar.

Expansion of a long coding repeat in the prion protein gene

Several other repeating sequences with units much longer than 3 bp have been found to be associated with inherited disease and most of their stories go back before the start of this project. Human prion diseases (Creutzfeldt-Jakob disease, Gerstmann-Sträussler syndrome *etc.*) occur in inherited, sporadic and acquired forms. The inherited forms are associated with coding mutations in the prion protein gene. Identification of one of these mutations allows definitive diagnosis and has resulted in a widening of the previously-recognized clinical spectrum (Collinge & Palmer, 1994). At least 11 pathogenic point mutations - causing amino-acid substitutions, or in one case a stop codon - have been reported and all the other pathogenic mutations are expansions in a 24 bp repeat encoding an octapeptide repeat in the protein.

All the expansions reported to March 1994 are listed by Goldfarb *et al.* (1994). The normal gene contains 5 copies of the repeat with four different variants at the nucleotide level, named R1 - R4. R1 has in fact 27 bp encoding 9 amino-acids. The other three repeats encode the same 8 amino-acids (numbers 1, 2 and 4 - 9 of R1) but differ at third-base positions. The normal sequence is R1, R2, R2, R3, R4. Some of the expanded alleles contain new variants amongst the extra copies, again not changing the amino-acid sequence of the repeat. All the expanded repeats have one copy of R1 at the beginning and one copy of R4 at the end so the additional copies are all extra copies of R2 or R3 or new variants. *E.g.* an allele with 9 extra copies was found to be R1, R2, R2, R3, R2, R3g, R2a, R2, R2, R2, R3g, R2, R3, R4, where R2a and R3g are new variants. Seven different expanded alleles were reported by Goldfarb *et al.* (1994), containing 2, 4, 5, 6, 7, 8, and 9 extra repeat units and all their bearers had had brain disease except for the one with 4 extra copies who had died of cirrhosis aged 63. However, Collinge & Palmer (1994) had seen an unpublished case

of classical CJD with 4 extra repeats albeit with a slightly different base sequence but coding for the same amino-acids. This was later published (Campbell *et al.*, 1996).

Since these reviews, several other reports of additional expanded repeat alleles have been published (Oda *et al.*, 1995; Krasemann *et al.*, 1995; Capellari *et al.*, 1997) but none has more than a total of 14 repeat units and I have not seen a report of change in number of repeats within a family. Looking at the potential for DNA secondary-structure formation, it is notable that the coding strand is about 50% G residues - *e.g.* the coding strand sequence of R2 is CCT CAT GGT GGT GGC TGG GGG CAG - and therefore might be liable to form tetraplexes (see Chapter 5).

Minisatellite repeats that regulate gene expression

Krontiris *et al.* (1993) reported that possession of any one of a group of rare alleles of a minisatellite 1 kb downstream of the *HRAS1* (*H-RAS1*, c-Ha-*ras*-1) proto-oncogene was a major risk factor for various common types of cancer. The minisatellite consists of about 30 - 100 repeats of a 28-bp consensus sequence, d(CACTCCCCCTTCTCTCCAGGGGACGCCA)·d(TGGCGTCCCCTGGAGAG AAGGGGGAGTG) (Capon *et al.*, 1983). Four common alleles account for 94% of all alleles in Europeans and there are more than two dozen other known alleles. Krontiris *et al.* (1993) found that these rare alleles were much more common in cancer patients and that their possession accounted for about 1 in 11 cancers of the breast, colorectum and bladder. The risk to people with two rare alleles seemed to be at least twice that for people with only one rare allele. The authors said that though they could not exclude the possibility that the rare alleles were simply linkage markers for some other cause of pathogenesis, they proposed the alternative that the rare minisatellite alleles disrupt the controlled expression of nearby genes including *HRAS1*. This was because they had previously shown that the minisatellite binds at least four members of the *rel*/NF-κB family of transcriptional regulators (Trepicchio & Krontiris, 1992) and they had shown allele-specific effects of the *HRAS1* minisatellite on reporter-gene activation (Green & Krontiris, 1993). The *rel*/NF-κB

proteins include positive and negative regulators of transcription and Krontiris *et al.* (1993) proposed that different ones bound to different variant repeats and thus the balance of suppression and activation could be different with different minisatellite alleles.

Trepicchio & Krontiris (1993) showed that a minisatellite in the intron separating the diversity and joining regions of the human immunoglobulin heavy-chain gene binds a transcriptional regulatory protein closely related to members of the myc/HLH family of proteins and that this binding appeared to sequester the factor in a form that could no longer activate transcription in a model system. The whole minisatellite is deleted when the IgH gene is rearranged in B lymphocytes, raising the possibility of stage-specific regulation. The repeating unit is 50 bp long and four common and two rare larger alleles were found originally with ?8, 10 12, 16, 20 and 32 repeat units in 93 people (186 alleles, Silva *et al.*, 1987). It has yet to be shown whether there are phenotypic differences in people with different genotypes at this minisatellite.

365 bp 5' to the insulin gene, and in its promoter region, is a minisatellite (or VNTR) of 14 - 15 bp repeats of consensus sequence d(ACAGGGGT$^{G}/_{C}{}^{T}/_{C}$GGGG)· d(CCCC$^{A}/_{G}{}^{C}/_{G}$ACCCCTGT) (Bell *et al.*, 1982). Repeat numbers vary from about 30 to >540 (Rotwein *et al.*, 1986) but are divided into three distinct size classes, I - III (Bell *et al.*, 1984), of average size 40, 85 and 157 repeats respectively (Kennedy *et al.*, 1995) and 14 subunit variants have been described, numbered 'a' - 'n' (Rotwein *et al.*, 1986). In Caucasoids, homozygosity for class I alleles was found to be associated with type 1 diabetes (insulin-dependent diabetes mellitus, IDDM) (refs in Bennett *et al.*, 1995; Kennedy *et al.*, 1995). Takeda *et al.* (1989) showed negative regulation of the insulin gene by one of the VNTR alleles and Hammond-Kosack *et al.* (1992a; b; c; 1993) showed that the G-rich strand of the VNTR could adopt a variety of quadruplex configurations *in vitro* and *in vivo* and even when wound on histones.

Then in two adjacent papers in 1995, evidence of a role for the VNTR became more concrete. Bennet *et al.* showed that susceptibility to type 1 diabetes at the

locus *IDDM2* resided in the VNTR. In general class I alleles conferred a greater susceptibility to diabetes than class III alleles except for one particular class I allele, containing 'e' variant repeat units, that was protective. There also appeared to be a parent-of-origin effect suggesting imprinting. By exploiting polymorphisms in the 3'-untranslated region they were able to distinguish RNA from the two alleles from seven human pancreata and in each case found different levels of expression by the two alleles *in vivo*. In the other paper, Kennedy *et al.* showed in *in vitro* studies that the VNTR binds a transcription factor, Pur-1, and activates transcription from a linked downstream promoter. They also found that Pur-1 binds with different affinity to different variant repeats but that in general long alleles confer greater activity to the insulin gene than do short alleles. This particular case seems to be an exception to the general rule that longer repeat tracts carry more disease risk. The mutation rates of all the minisatellite loci mentioned in this section so far remain to be established.

Progressive myoclonus epilepsy type 1 (EPM1, Unverricht-Lundborg disease) is an autosomal recessive disorder due to loss of function of the gene, *CST6* (or *STFB* or *CSTB*), encoding cystatin B (*alias* stefin B), a cysteine protease inhibitor. Lalioti *et al.* (1997a) found no mutation in the exons or intron-exon junctions of 50/58 alleles, though mRNA levels were markedly reduced, and one of the possibilities they considered was a mutation in a regulatory region.

Lafrenière *et al.* (1997) found that the most common mutation was an expansion of ~600 - 900 bp in the 5'-flanking region (*i.e.* 5' of the transcription start site). They found it was within a region that consisted of two or three copies of a 12 bp repeat unit d(CCCCGCCCCGCG)·d(CGCGGGGCGGGG) on control chromosomes but were unable to amplify the expanded region by PCR.

Virtaneva *et al.* (1997) found that an allele containing an expansion of about 1 kb shrank to 0·4 kb on cloning in λ and shrank further to 0·2 kb when subcloned in a plasmid. After refining their PCR and sequencing conditions they were still not able to sequence right across the λ clone but concluded from their gels that at the 5' end of

the expanded region, following two copies of the 12 bp repeat, there was an unknown number of copies of an 18 bp repeat, d(CCCTCGCCTCTCAGTCTG)·d(CAGACT GAGAGGCGAGGG), and at the other end of the expanded region there were copies of a 15 bp repeat, d(CTCCTCGCCCACGAG)·d(CTCGTGGGCGAGGAG) before a final single copy of the 12-mer repeat. They suggested that the 15-mer could have arisen by mutation from the 12-mer and that after some expansion in numbers, the 18-mer could have arisen from a copy of the 15-mer.

Neither Lafrenière *et al.* (1997) nor Virtaneva *et al.* (1997) found any instance of change in size of the expanded region between generations but the former group noted that different lengths were seen on chromosomes with the same haplotype, showing that size changes had occurred over longer periods of time. Virtaneva *et al.* (1997) found expanded alleles, ranging from 0·5 - >1·5 kb, in 1 in 48 random Finnish controls, compatible with an estimated carrier rate of 1:50 - 1:80 in Finland. Lafrenière *et al.* (1997) speculated that the expansion might reduce expression of the gene by disrupting the promoter or by becoming hypermethylated.

The week after the paper of Virtaneva *et al.* (1997) was published, Lalioti *et al.* (1997b) published their conclusion of the expansion of perfect repeats of the 12-mer to >60 copies in pathogenic alleles. They too could not demonstrate instability of these large alleles but in screens of normal families from several populations they found a few alleles in the 12 - 17 repeats range and these were transmitted unstably in 29% of transmissions, the largest change being 13 → 17 units, but only when transmitted by males. Mandel (1997) has made the interesting observation that in terms of base-pairs, rather than numbers of repeats, these alleles fall into the same unstable size range as trinucleotide repeats. Homozygotes for alleles in this range were unaffected. No increase in methylation was detected in the enlarged alleles (Lalioti *et al.*, 1997b), indicating that suppression of expression was achieved by some other means, presumably by affecting DNA-binding factors.

Subsequently Lalioti *et al.* (1997c) concluded that the 15-mer and 18-mer of Virtaneva *et al.* (1997) were sequencing artefacts. They pointed out that the 15-mer

would be cleaved by *Sac*I (recognition site GAGCTC) and that the 18-mer would be cleaved by *Dde*I (recognition site CTCAG) and showed that restriction fragments corresponding to cleavage within the repeat did not occur. More recently (Lalioti *et al.*, 1998) they have found a range of ~30 - ~75 repeats in patients and have observed changes in size within families.

The level of mRNA from expanded alleles was reduced in blood leukocytes but was normal in lymphoblastoid cell lines (Lalioti *et al.*, 1997b) and one of the possibilities that the authors discussed was that in certain cells the *CSTB* gene might use an earlier transcription start site, *i.e.* this repeat too may be transcribed.

Other possible effects of expanded repeats on disease susceptibility

Krontiris *et al.* (1993) suggested that the findings on the *HRAS1*, IgH and IDDM2 minisatellites, and the discovery of binding sites for SP1 in the intronic minisatellite of the interleukin-1α gene suggested the possibility of a broader contribution of minisatellites to the genetic risk of disease. O'Donovan *et al.* (1996b) had a similar thought regarding trinucleotide repeats. They had found (O'Donovan *et al.*, 1996a) an association of expanded d(CAG)·d(CTG) repeats with schizophrenia, and with bipolar affective disorder, that could not alone account for the inheritance or for the apparent anticipation (the average size of d(CAG)·d(CTG) repeat tracts detected by RED was significantly larger in both groups than in controls but there was complete overlap and no locus was identified). O'Donovan *et al.* (1996b) tested the hypothesis that expanded trinucleotide repeats might confer increased risk of common disorders with multifactorial aetiology. If this were so then, amongst people selected for being healthy, older people should have shorter average repeat tract lengths than younger people. The authors used RED to measure maximum d(CAG)·d(CTG) repeat sizes in blood donors and found a significant negative correlation with age. This did not occur in populations unselected for health. Whether this latter finding will be sustained by future work remains to be seen but

the importance of repeated sequences in promoting genomic rearrangements and in affecting fitness, at least at a growing number of individual loci, is clear.

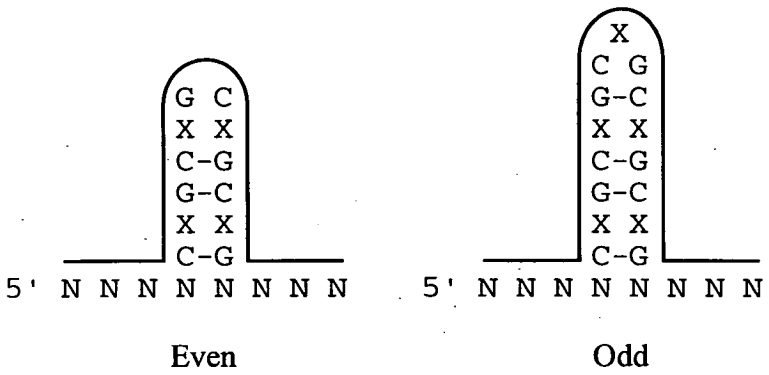
The work of this project

In the work to be described, I have used bacteriophage λ derivatives containing a long palindrome to study behaviour of trinucleotide repeats *in vivo*. In double-stranded DNA palindromes, *i.e.* inverted repeats, opposite halves of the palindrome on the *same* strand are complementary to one-another and, may anneal to form a hairpin, or a cruciform if both strands so pair. Long DNA palindromes are not recovered in DNA libraries when they are introduced into bacteria. Either they are wholly or partially deleted (instability) or they cause failure of replication of the vector (inviability) (reviewed in Leach, 1994). The threshold for this inviability is about 150 - 200 bp total length of the palindrome. Further investigations led to the discovery that mutations in *E. coli* genes *sbcC* and *sbcD* allow the propagation of bacteriophage λ derivatives with long palindromes (Chalker *et al.*, 1988; Gibson *et al.*, 1992). However, in *sbcC* mutant hosts, inviability is not totally overcome and the plaque size of palindrome-containing phage is acutely sensitive to the central sequence of the palindrome (Davison & Leach, 1994a).

Davison & Leach (1994a) showed that central sequences predicted to stabilise DNA hairpins reduce plaque size. In all positions outside the central two base-pairs of a perfect palindrome, C and G produced smaller plaques than A and T. This is the reverse of what would be expected if melting were the rate-limiting step, as it appears to be *in vitro*. The observed effect diminishes with distance from the centre, suggesting that formation of the first few intra-strand base pairs ('protocruciform' formation) is the rate limiting step *in vivo*. Also, sequences known to adopt two-base loops *in vitro* generate smaller plaques than sequences known to adopt four-base loops *in vitro* (Davison & Leach, 1994b). These studies showed that there is no correlation between predicted central melting and plaque size, but that hairpin-loop

stability correlates inversely with plaque size. It was therefore concluded that the measurement of plaque areas is a reliable assay for the stability of DNA hairpin-loops (Davison & Leach, 1994a,b).

Thus trinucleotide repeats, or DNA of any other sequence, can be inserted into the centre of a palindrome and, if the sequence forms a stable hairpin with a fold that corresponds to the centre of the palindrome, plaque size will be small, but if not, the plaque size will increase. It was reasoned that if d(CXG)·d(CX'G) repeats form imperfect hairpins held together by C·G pairing, as proposed by Sinden and Wells (1992), these hairpins would have the potential to exist in two possible forms comprising odd or even numbers of repeat units that differ only in the nature of the loop formed at the apex of the hairpin (Leach, 1994) as shown below.



Therefore, by construction of a series of 'phage with increasing numbers of trinucleotides in the centre of a palindrome it could be observed whether one or other of these loops might be particularly stable *in vivo*. Chapter 3 describes development of the plaque assay. Chapter 4 gives the results with λ phages containing different numbers of d(CAG)·d(CTG) repeats inserted into the centre of a long palindrome. These are compared with results with odd and even numbers of d(GAC)·d(GTC) repeats and sequences known to form two-base and four-base loops as well as 'phage constructed to test the effect of sequence context on the results.

With d(CGG)·d(CCG) repeat DNA different investigators came to quite different conclusions about what secondary structures are formed by the single

strands *in vitro*. Chapter 5 analyzes these papers to determine the reasons for the differences and what are the most likely structures to form *in vivo*. Then Chapter 6 gives the results of plaque assays with series of d(CGG)·d(CCG) repeats in different frames relative to the centre of the palindrome.

Chapter 7 describes the construction and use of a bacteriophage to test whether the orientation of the repeat sequences relative to the origin of replication affects the results. The effect of having a single base-pair of asymmetry in the palindrome is also tested because this affects whether the palindrome can be cleaved on one side of the centre only so that the central sequence can be verified. (Secondary structure prevents sequencing across a long palindrome if it is not cleaved but with cleavage near the centre on both sides the central sequence will be lost.) Then results on a series of d(GAA)·d(TTC) repeats is given.

Three publications have emerged from this work so far (Darlow & Leach, 1995; 1998a,b) and most chapters contain material from these, modified or unmodified. Copies of these publications are appended to this thesis.

Chapter 2

Materials and Methods I

Materials

Bacteria - *Escherichia coli* strains

The bacteriophage used in this project had long DNA palindromes and could not be propagated on wild-type or any rec^+ *E. coli* and the following strains were used accordingly:

JC9387: F⁻ *thi*-1 *his*-4 Δ(*gpt-proA*)62 *argE*3 *thr*-1 *leuB*6 *kdgK*51 *rfbD*1 (?) *ara*-14 *lacY*1 *galK*2 *xyl*-5 *mtl*-1 *tsx*-33 *rpsL*31 *recB*21 *recC*22 *sbcB*15 *sbcC*201 ***sup*^{0}**. This was derived from AB1157 (Howard-Flanders & Theriot, 1966) in the laboratory of A.J. Clark and obtained from F. Stahl. The strain was used for general management of the 'phage.

R594: F⁻ *lac galK*2 *galT*22 *rpsL*179 (Str^r) (Campbell, 1965). This strain is rec^+ and was used for detection of 'phage which had lost their palindromes.

N2364: F⁻ *thi*-1 *his*-4 Δ(*gpt-proA*)62 *argE*3 *thr*-1 *leuB*6 *kdgK*51 *rfbD*1 (?) *ara*-14 *lacY*1 *galK*2 *xyl*-5 *mtl*-1 *tsx*-33 *supE*44 *rpsL*31 *sbcC*201 *phoR*79::TN*10*. This was derived from AB1157 (Howard-Flanders & Theriot, 1966) by Lloyd & Buckman (1985) and was obtained from R.G. Lloyd. 'Phage with long palindromes would grow on this strain, but not as well as on JC9387. Its principal use was in the plaque assay (see Chapter 3) as it gave better discrimination between 'phage with different central sequences in their palindromes with respect to plaque size than did JC9387.

(**BHB2688:** *recA* (λ*imm*^{434} *cIts b*2 *red*3 *Eam*4 *Sam*7)/λ and **BHB2690:** *recA* (λ*imm*^{434} *cIts b*2 *red*3 *Dam*15 *Sam*7)/λ (Hohn, 1979), obtained from N. Murray, were used for preparation of bacteriophage packaging extracts.)

Bacteriophage λ strains

DRL152: This was the only strain used that did not have a long palindrome and was used for comparison of its plaque size. It is *spi*6, *cI*857, χ^+C153.

*spi*6 is a deletion of λ rendering it *red gam*. *cI*857 is a temperature sensitive repressor. $\lambda\chi^+C153$ phage have a χ^+ site at bp 38481 - 38488 enabling efficient replication of *red gam* phage.

DRL167: This was the parent bacteriophage from which all other bacteriophage used in this work were constructed. Its description is *pal*462(*Sac*I), *spi*6, *cI*857, χ^+C153.

*pal*462(*Sac*I) indicates that the 'phage contains (near the centre of the chromosome) a 462 bp perfect palindrome (sequence in Appendix 1) with a *Sac*I site at the centre. There is no other *Sac*I site in the 'phage. The palindrome has *Eco*RI sites at the ends and is the same one that is present in DRL133 (Chalker *et al.*, 1993), from which DRL167 was derived by a cross with DRL152 (Davison & Leach, 1994a). The latter introduced the Chi site which rendered plaques large enough to be measured on an image analyser. The construction of the palindrome is best described in Allers (1993).

DRL176: This 'phage was used in this work as a plaque-size reference. It is the same as DRL167 except that it has *pal*476(*Bam*H1). It was constructed from DRL167 by ligating the following insert into the *Sac*I site:

```
5'        GTGGATCCACAGCT  3'
3'  TCGACACCTAGGTG        5'
```

The *Bam*H1 site is in bold type. The 'phage was constructed by Allers (1993).

DRL199 and **DRL207**, used in one study, were similarly derived from DRL167 by A. Davison and their central sequences are given in Table 4.1, Chapter 4.

Other 'phage used were constructed in this work, as described under Methods I, below, and in chapters 4, 6 and 7, and are listed in Tables 2.1 and 2.2.

Table 2.1 Bacteriophage strains derived directly from DRL167

Insert	Working Name		Assigned
	Expt.	Isolate	Name
5' GGTCTCG(CAG)$_1$CGAGACCAGCT 3' 3' TCGACCAGAGC(GTC)$_1$GCTCTGG 5'	1 (and 4	12 8)	DRL220
5' GGTCTCG(GAC)$_1$CGAGACCAGCT 3' 3' TCGACCAGAGC(CTG)$_1$GCTCTGG 5'	2	1 (9)	DRL225
5' GGTCTCG(CAG)$_2$CGAGACCAGCT 3' 3' TCGACCAGAGC(GTC)$_2$GCTCTGG 5'	5	26	DRL221
5' GGTCTCG(GAC)$_2$CGAGACCAGCT 3' 3' TCGACCAGAGC(CTG)$_2$GCTCTGG 5'	6	20 (21)	DRL226
5' GGTCTCG(CAG)$_3$CGAGACCAGCT 3' 3' TCGACCAGAGC(GTC)$_3$GCTCTGG 5'	7	25	DRL222
5' GGTCTCG(GAC)$_3$CGAGACCAGCT 3' 3' TCGACCAGAGC(CTG)$_3$GCTCTGG 5'	8	15	DRL227
5' GGTCTCG(CAG)$_4$CGAGACCAGCT 3' 3' TCGACCAGAGC(GTC)$_4$GCTCTGG 5'	9	3	DRL223
5' GGTCTCG(GAC)$_4$CGAGACCAGCT 3' 3' TCGACCAGAGC(CTG)$_4$GCTCTGG 5'	10	3 (4)	DRL228
5' GGTCTCG(CAG)$_5$CGAGACCAGCT 3' 3' TCGACCAGAGC(GTC)$_5$GCTCTGG 5'	11	2	DRL224
5' GGTCTCG(GAC)$_5$CGAGACCAGCT 3' 3' TCGACCAGAGC(CTG)$_5$GCTCTGG 5'	12	3	DRL229
5' GGTCTCG(CCG)$_1$CGAGACCAGCT 3' 3' TCGACCAGAGC(GGC)$_1$GCTCTGG 5'	13	7, 32	DRL230
5' GGTCTCG(CCG)$_2$CGAGACCAGCT 3' 3' TCGACCAGAGC(GGC)$_2$GCTCTGG 5'	14	11, 24	DRL231
5' GGTCTCG(CCG)$_3$CGAGACCAGCT 3' 3' TCGACCAGAGC(GGC)$_3$GCTCTGG 5'	15	5, 10	DRL232
5' GGTCTCG(CCG)$_4$CGAGACCAGCT 3' 3' TCGACCAGAGC(GGC)$_4$GCTCTGG 5'	16	3, 12	DRL233
5' GGTCTCG(CCG)$_5$CGAGACCAGCT 3' 3' TCGACCAGAGC(GGC)$_5$GCTCTGG 5'	17	7, 8	DRL234
5' GGTCTCC(CAG)$_2$GGAGACCAGCT 3' 3' TCGACCAGAGG(GTC)$_2$CCTCTGG 5'	18	54	DRL237

5' GGTCTCC(GAC)$_2$GGAGACCAGCT 3' 3' TCGACCAGAGG(CTG)$_2$CCTCTGG 5'	19	2	DRL238
5' GGTCTCCACTTGTGGAGACCAGCT 3' 3' TCGACCAGAGGTGAACTCCTCTGG 5'	20	9	DRL235
5' GGTCTCCAGTTCTGGAGACCAGCT 3' 3' TCGACCAGAGGACAAGTCCTCTGG 5'	21	1	DRL236
5' GGTCTCG(GCC)$_1$CGAGACCAGCT 3' 3' TCGACCAGAGC(CGG)$_1$GCTCTGG 5'	22	2, 7	DRL247
5' GGTCTCG(GCC)$_2$CGAGACCAGCT 3' 3' TCGACCAGAGC(CGG)$_2$GCTCTGG 5'	23	1, 9	DRL248
5' GGTCTCG(GCC)$_3$CGAGACCAGCT 3' 3' TCGACCAGAGC(CGG)$_3$GCTCTGG 5'	24	3, 6	DRL249
5' GGTCTCG(GCC)$_4$CGAGACCAGCT 3' 3' TCGACCAGAGC(CGG)$_4$GCTCTGG 5'	25	8, 14	DRL250
5' GGTCTCG(GCC)$_5$CGAGACCAGCT 3' 3' TCGACCAGAGC(CGG)$_5$GCTCTGG 5'	26	33, 38	DRL251
5' GGTCTCG(CGC)$_1$CGAGACCAGCT 3' 3' TCGACCAGAGC(GCG)$_1$GCTCTGG 5'	27	29, 32	DRL252
5' GGTCTCG(CGC)$_2$CGAGACCAGCT 3' 3' TCGACCAGAGC(GCG)$_2$GCTCTGG 5'	28	34, 46	DRL253
5' GGTCTCG(CGC)$_3$CGAGACCAGCT 3' 3' TCGACCAGAGC(GCG)$_3$GCTCTGG 5'	29	50, 52	DRL254
5' GGTCTCG(CGC)$_4$CGAGACCAGCT 3' 3' TCGACCAGAGC(GCG)$_4$GCTCTGG 5'	30	11, 15	DRL255
5' GGTCTCG(CGC)$_5$CGAGACCAGCT 3' 3' TCGACCAGAGC(GCG)$_5$GCTCTGG 5'	31	13, 17	DRL256
L 5' GGCCTCGATGGCCCTGACTCGAGGCCAGCT TCGACCGGAGCTACCGGGACTGAGCTCCGG 5' **R**	32	2	DRL257
L 5' GGCCTCGAGTCAGGGCCATCGAGGCCAGCT TCGACCGGAGCTCAGTCCCGGTAGCTCCGG 5' **R**	32	1, 3, 4	DRL258

Strains listed were made by ligating the inserts into the *Sac*I site in the palindrome of DRL167. The insert could be orientated in either direction with respect to the bacteriophage genome and orientation could not be ascertained for any particular isolate because it was not possible to sequence across the palindrome. With the insert used in construct 32 however, orientation could be ascertained (see Chapter 7).

Table 2.2 Bacteriophage strains derived from DRL257 (orientation A) and DRL258 (orientation B)

Parent 'phage	Insert	S/H	Working Name Expt.	Isolate	Assigned Name
A	5' GCTCG(CAG)$_3$CGAGC GCTCGAGC(GTC)$_3$GCTCGAGCT	3sA	35	4	DRL261
A	5' ACTCG(CAG)$_1$CGAGC GCTTGAGC(GTC)$_1$GCTCGAGCT	1aA	36	8, 14	DRL262
A	5' ACTCG(CAG)$_2$CGAGC GCTTGAGC(GTC)$_2$GCTCGAGCT	2aA	37	7	DRL263
A	5' ACTCG(CAG)$_3$CGAGC GCTTGAGC(GTC)$_3$GCTCGAGCT	3aA	38	12 (20)	DRL264
B	5' GCTCG(CAG)$_1$CGAGC GCTCGAGC(GTC)$_1$GCTCGAGCT	1sB	39	9	DRL265
B	Same as 35	3sB	41	29 (34)	DRL267
B	Same as 36	1aB	44	1, 7	DRL270
B	Same as 37	2aB	45	2	DRL271
B	Same as 38	3aB	46	4 (19)	DRL272
B	5' GCTCG(TTC)$_1$CGAGC GCTCGAGC(AAG)$_1$GCTCGAGCT		47	3	DRL273
B	5' GCTCG(TTC)$_2$CGAGC GCTCGAGC(AAG)$_2$GCTCGAGCT		48	2 (17)	DRL274
B	5' GCTCG(TTC)$_3$CGAGC GCTCGAGC(AAG)$_3$GCTCGAGCT		49	6 (9)	DRL275
B	5' GCTCG(TTC)$_4$CGAGC GCTCGAGC(AAG)$_4$GCTCGAGCT		50	15	DRL276
B	5' GCTCG(TTC)$_5$CGAGC GCTCGAGC(AAG)$_5$GCTCGAGCT		51	(5) 11	DRL277

S/H: shorthand name. For details of construction see under Methods I and Chapter 7.

Oligonucleotides

Oligonucleotides were supplied by Oswel DNA Service (University of Edinburgh, later transferred to University of Southampton), and later by Perkin Elmer (Warrington) or Genosys (Cambridge), whichever gave the best price.

Media

The following quantities are for final volumes of 1 litre unless otherwise stated, made up with distilled water and autoclaved for 20 min at 15 lb/in^2.

L broth: 10 g Bacto tryptone, 5 g yeast extract, 10 g NaCl

Lpc (for plating cells)**:** To a 100 ml bottle of L broth were added 2 ml 20% maltose, 1 ml 1 M $MgSO_4$ and 200 µl Vitamin B_1 (5 mg/ml). This could not be reautoclaved and was kept on the bench at room temperature where it could easily be seen whether there was any contaminant growth.

L agar: As L broth but with 15 g Oxoid® agar.

L bottom agarose (1·2%): This was used for production of plate-lysates of bacteriophage to avoid the possibility of constituents of agar inhibiting restriction enzymes used with DNA extracted from the 'phage. On the day before plate lysis 6 g agarose was added to a 500 ml bottle of L broth which was then reautoclaved and the contents mixed by inverting several times. The bottle was then kept at 60°C overnight. In the morning the bottle was cooled to 46°C in a water-bath and the following additions made: 7·5 ml 20% glucose, 5 ml 1 M Tris (pH 7·5), 1 ml 1 M $MgSO_4$, 800 µl 0·5 M $CaCl_2$, 200 µl 10 mM $FeCl_3$ and 1 ml vitamin B_1 (5 mg/ml).

L top agarose (0·4%): As for L bottom agarose but 0·4 g agarose was added to a 100 ml bottle of L broth and no additions were made.

BBL bottom agar: 10 g trypticase (Becton-Dickinson), 5 g NaCl, 10 g Bacto-agar (Difco). (BBL stands for Baltimore Biological Laboratories.)

BBL top agar: As BBL bottom agar but only 6·5 g Bacto-agar.

PSQ agars: Quantities:

bottom agar: Per 500 ml bottle: Original quantities, as used for work described in Chapter 4: 5 g Difco agar, 5 g Select Trypticase, 7 g NaCl, 5 ml 1 M Tris.HCl, 495 ml water. Increased quantities used for work described in later chapters: 5·25 g Difco agar, 5·25 g Select Trypticase, 7·35 g NaCl, 5·25 ml 1 M Tris.HCl, 519·75 ml water.

top agar: Per 100 ml bottle: 0·5 g Difco agar, 1 g Select trypticase, 1·4 g NaCl, 1 ml 1 M Tris.HCl, 99 ml water.

Method: The agar was put into the bottle. The other ingredients were mixed in a 600 ml or 150 ml beaker as appropriate and added to the bottle. After autoclaving the bottle was inverted several times to mix the melted agar lying at the bottom.

'Phage buffer: 3 g KH_2PO_4, 7 g Na_2HPO_4, 1 mM $MgSO_4$, 1 mM $CaCl_2$, 1 ml gelatin (1% w/v)

Stock solutions

0·5 M $CaCl_2$: made up with distilled water and autoclaved.

CsCl: For 1·3 g/ml solution, 31·24 g was weighed out, for 1·5 g/ml, 45·41 g and for 1·7 g/ml, 56·24 g, 'phage buffer was added to bring each total weight to 100 g, CsCl dissolved and left overnight before use for precipitated gelatin to settle.

0·5 M EDTA: made up with distilled water, adjusted to pH 8 with NaOH and autoclaved.

10 mM $FeCl_3$: Concentration only approximate because the compound absorbs atmospheric water so rapidly that it is impossible to weigh accurately.

20% Glucose: made up in distilled water and filter-sterilized.

20% Maltose: made up in distilled water and filter-sterilized.

100 mM β-mercaptoethanol: 1 µl β-mercaptoethanol + 142 µl distilled water.

1 M $MgSO_4$: made up with distilled water and autoclaved.

3 M NaAc (sodium acetate): made up as in Sambrook *et al.* (1989), p. B13 but adjusted to pH 5·3, not 5·2, autoclaved.

5 M NaCl: made up with distilled water and autoclaved.

30% PEG 6000, 3M NaCl: made up with distilled water and stored at 4°C.

1 M Piperidine: 85 µl piperidine + 915 µl distilled water or other amounts in the same proportion.

Sanger dye/Stop buffer/Stop solution/Sequencing gel sample loading buffer: 95% formamide, 20 mM EDTA, 0·05% bromophenol blue and 0·05% xylene cyanol FF.

20 × TAE: 96·8 g/l Tris, 22·85 ml glacial HAc per litre.

10 × TBE: 108 g Tris, 55 g boric acid, 40 ml EDTA (0·5 M, pH 8) per litre.

5 × TAE gel-loading buffer: 5 ml 20 × TAE, 20 mg bromophenol blue (final concentration 0·2%), 10 ml 0·5 M EDTA, pH 8 (final conc. 0·25 M) 3 g Ficol 400 (final conc. 15%) made up to 20 ml with distilled water.

1 M Tris: adjusted to pH 7·5 with HCl.

10 × TE: 100 mM Tris (12·11 g/l), 10 mM EDTA (3·72 g/l), adjusted to pH 7·5 with HCl and autoclaved.

10 × TM: 100 mM Tris (12·11 g/l), 100 mM $MgSO_4$ (24·65 g/l), adjusted to pH 7·5 with HCl and autoclaved.

Vitamin B_1 (thiamine): 5 mg/ml made up in about 20 ml distilled water, Millipore filtered and stored in a foil wrapped container at 4°C.

Organic liquids

Chloroform-isoamylalcohol: 24 volumes of chloroform were mixed with 1 volume of isoamylalcohol and kept in a dark bottle.

'70%' Ethanol: 7 volumes ethanol: 3 volumes distilled water.

Tris-saturated phenol and phenol-chloroform: Distilled, liquefied (88%) phenol (Rathburn Chemicals) was stored in 50 ml volumes at –20°C, protected from light in foil-wrapped polypropylene tubes. On thawing one tube, half of the phenol would be decanted into a similar tube and 25 mg of 8-hydroxyquinoline added to each (*i.e.* 0·1% w/v), followed by 25 ml 10 × TE. The tubes were then shaken vigorously to emulsify the liquids and put on a vertical wheel mixer for 10 min., then centrifuged at 4·5 krpm for 5 min.. The aqueous layer was removed from each and equilibration repeated twice more but with 1 × TE and the aqueous layer removed again. Then, if making phenol-chloroform, an equal volume (usually less than 25 ml by this stage) of chloroform-isoamylalcohol would be added to the phenol, the mixture centrifuged as before, and a further aqueous layer removed. Then the phenol or phenol-chloroform was carefully layered with 2 ml 1 × TE containing 4 µl (*i.e.* 0·2% v/v) β-mercaptoethanol. The tubes were then wrapped in foil and stored at –20°C.

Enzymes

DNase and RNase: Quantities:

DNase I (10 mg/ml): 10 mg DNase I, 50 µl 1 M Tris (pH 7·5), 100 µl 100 mM NaCl, 10 µl 10 mg/ml BSA, 1 µl 1 M DTT, 500 µl glycerol, 340 µl distilled water. Total, 1 ml.

RNase A (10 mg/ml): 10 mg RNase A, 10 µl 1 M Tris (pH 7·5), 150 µl 100 mM NaCl, 840 µl distilled water.

Treatment: Heat to 100°C for 10 - 15 min (floating tubes in a beaker of boiling water) and allow to cool slowly to room temp.. Store at –20°C.

Pronase (serine protease from *Streptomyces griseus*): 20 mg pronase, 10 µl 1 M Tris (pH 7·5) 100 µl 100 mM NaCl, distilled water 870 µl, total 1 ml, incubated at 37°C for 1 hr to autodigest and stored at –20°C.

Polynucleotide kinase: was obtained from Sandra Bruce (Institute of Cell & Molecular Biology, University of Edinburgh). The dilution buffer was 25 mM Tris (pH 7·5), 10 mM β-mercaptoethanol, 50% glycerol made up with sterile distilled water and vortexed to mix. 1/25 dilution renders the concentration about 1 unit/µl. Commercial T4 DNA ligase buffer was used as the reaction buffer.

Restriction enzymes, T4 DNA ligase, Klenow fragment and heat-stable polymerases for PCR were obtained from various companies and used with the manufacturers' buffers and according to their instructions except that restriction enzymes were usually used at somewhat higher than recommended concentrations and were usually incubated for at least 2 hr.

Gel solutions

Agarose: For agarose gel electrophoresis Flowgen Routine Electrophoresis Grade agarose was used, made to the desired percentage w/v with 1 × TAE, melted in a microwave oven and used immediately as described under Methods.

Polyacrylamide: For comparing sizes of labelled palindromes of around 500 bp to detect differences of as little as 3 bp (one trinucleotide in the central insert) 5% denaturing polyacrylamide gels were used: 42·5 g urea, 10 ml 10 × TBE and 12·5 ml 40% w/v acrylamide/bisacrylamide - 2·105 w/v (ratio 19:1, Scotlab, Luton, Beds) made up to 100 ml. This was sufficient for two gels and could be kept at 4°C for two weeks.

For Maxam-Gilbert sequencing of oligonucleotides 12% polyacrylamide was used: the same recipe but 30 ml of the acrylamide/bisacrylamide.

Methods I

Routine bacterial and 'phage culture

Long-term stocks of bacteria were prepared by adding 5 drops of sterile 100% glycerol to 1 ml of overnight culture in a 1·5 ml Eppendorf tube which was then sealed with Parafilm and stored at –70°C.

Plate cultures (*i.e.* single colony cultures) were prepared by streaking bacteria from long-term stocks to single colonies on L agar plates and incubating overnight at 37°C. The plates were then kept at 4°C.

'Overnight cultures' (*i.e.* small liquid cultures) were prepared by inoculating 5 ml of L broth in a ½ oz bijou bottle with one colony from a plate culture and incubating on a rocker at 37°C overnight. The cultures were kept at 4°C.

Plating cells: 0·5 ml of overnight culture was added to 4·5 ml of Lpc in a sterile ½ oz bijou bottle and incubated on a rocker at 37°C to about mid-log phase (2 hr for R594, 2 hr 10 min for JC9387 and 2 hr 15 min for N2364). The culture was then diluted with 5 ml 1 × TM. The cells were kept at 4°C for up to a week.

'Phage suspensions of various titres were made, as appropriate to the protocols to be described, by diluting plate lysates with 'phage buffer, or by picking a plaque from a plate with a sterile Pasteur pipette and putting it into 1 ml (or a smaller volume for small plaques) of 'phage buffer and leaving it at least 1 hr at room temperature or at least 2 hr at 4°C for the phage to diffuse out of the agar, or by stabbing a plaque with a sterile wooden toothpick and leaving it to stand in 1 ml of 'phage buffer for the same times. The use of a Pasteur pipette to pick plaques is more time-consuming but yields about an order of magnitude higher concentration.

'Phage plating: To plate 'phage, 250 ml of plating cells were put into a sterile 12 × 75 mm glass tube (with a metal cap) and a quantity of 'phage suspension added depending upon the titre of the suspension and the density of plaques required. The

tube was then left for the 'phage to adsorb to the cell surfaces. (I usually used 15 - 20 min at 37°C because this is what I was taught at the beginning of the project but 10 min at room temperature is sufficient.) In the meantime, top agar (or agarose), kept molten in a 60°C oven was cooled to about 46°C in a water-bath. Then 2·5 ml of top agar (or agarose) was added to the tube and the contents of the tube immediately poured onto set bottom agar (or agarose) in a previously labelled 'plate' (*i.e.* Petri dish). The plate was then tipped in a rotary manner to spread the top agar/agarose before it set and then left on a level surface with the lid on for 5 min to set and then incubated right-way-up at 37°C.

For preparation of 'phage by plate lysate, L agarose plates poured within the hour were used (see under 'Cloning oligonucleotide inserts in the palindrome centre of bacteriophage λ DRL167'). For plating 'phage to pick or count plaques BBL agar was used rather than L agar because L agar is too rich causing the uninfected cells to grow very rapidly, resulting in minute plaques. It was found best to leave plates at least three days to dry at room temperature otherwise plaques could be very large (5 mm or more in diameter) and were often confluent. In emergencies, younger plates were dried in a plate-drying oven but this always resulted in an uneven thickness of bacterial lawn across the plate and a marked gradient of plaque size and plaque density across the plate, even if the plates were frequently turned in the oven. Such plates could be used for picking plaques but were not useful for titring. PSQ plates left over form plaque size quantification were satisfactory for both purposes.

Titring 'phage suspensions: This was done by plating known volumes of known dilutions of the suspensions and counting the plaques. For a fresh plate lysate 4 tubes were labelled 10^{-2}, 10^{-4}, 10^{-6} and 10^{-7}, to them 990, 990, 990 and 900 µl of 'phage buffer added respectively then serial dilutions made starting by adding 10 µl of the neat lysate to the first tube and so on to 100 µl from the third tube to the fourth. Then, if the titre of the lysate is about 5×10^{10} p.f.u./ml, plating 10 µl of the 10^{-7} dilution should give about 50 plaques and plating 100 µl of the 10^{-7} dilution (or

10 µl of the 10^{-6} dilution) should give about 500 plaques. However, some lysates may not be as concentrated and titre tends to drop by at least an order of magnitude as the lysate ages at 4°C, so this usually gave enough plaques for a reasonably accurate titre. Rarely, 100 µl of 10^{-6} dilution had to be plated.

DNA Purification

Phenol, phenol-chloroform and chloroform-isoamylalcohol extractions

All of the extractions were carried out in 1·5 ml Eppendorf tubes and in each case the volume of the organic liquid added was equal to that of the aqueous suspension. The tube was vortexed briefly and then centrifuged at full speed in a microcentrifuge (about 15 krpm) for 5 min for the phenol and phenol-chloroform extractions and 1 min for chloroform-isoamylalcohol extraction. The aqueous (upper) phase was then removed carefully and transferred to a clean Eppendorf tube for the next extraction or ethanol precipitation.

Ethanol and Isopropanol precipitations

For both precipitations a volume of 3 M NaAc was added to the DNA suspension equal to one ninth the latter's volume. Then, for ethanol precipitation, a volume of ethanol (kept at −20°C) equal to twice the total aqueous volume was added. This then gave the maximum volume from which DNA could be precipitated as about 450 µl (which + 50 µl NaAc + 1 ml ethanol = 1·5 ml). The tube was inverted several times to mix and incubated for 1 hr on ice or overnight at −20°C. Then it was centrifuged for 30 min at the ~15 krpm of the microcentrifuge at 4°C, the ethanol was removed, 70% ethanol (kept at −20°C) added to dissolve salt from the precipitate, and the tube recentrifuged for 15 min at 4°C. The supernatant was removed and the precipitate allowed to dry at room temperature with the lid of the

tube open and the tube either lying on its side or upright and near a Bunsen burner to create an updraft. This took up to about 15 min and then the DNA precipitate was ready for resuspension.

For isopropanol precipitation the procedure was similar but the isopropanol was kept at room temperature, the volume added was equal to 0·7 the volume of the aqueous suspension and incubation was for 10 - 15 min at room temperature (at which any RNA that might be present does not co-precipitate). Rinsing with 70% ethanol was the same.

Electrophoresis

Agarose gel electrophoresis

After melting the agarose (see Materials) in a conical flask, the flask was left on the bench to cool for a few minutes and then cooled to 46°C in a water-bath. In the meantime the gel tray was left at 4°C to cool. Gels were poured in horizontal 8-well, 2 × 3 in minigel or 14-well, 4·4 × 5·5 in 'midigel' electrophoresis apparati (Bethesda Research Laboratories). 5 × TAE loading buffer was added 1:4 to the samples before electrophoresis with 1 × TAE as the running buffer. 100 V was usually used for running unless the gel was to run overnight. Gels were stained with ethidium bromide (0·5 μg/ml) and usually destained in water before viewing on an ultraviolet light box and photographing with an electronic camera.

Polyacrylamide gel electrophoresis

Gels were poured in a Bio-Rad Sequi-Gen™ sequencing cell with a 40 cm front-plate. To seal the bottom of the cell, 60 μl of TEMED and 150 μl of 10% w/v ammonium persulphate were added to 12 ml of the required gel mixture (see Materials) - taking about 2 min to set - and for pouring the gel itself, 40 μl of

TEMED and 100 µl of 10% w/v ammonium persulphate were added to 35 ml of the gel mixture. A square-toothed comb was used. The running buffer was 1 × TBE. Gels were prerun at 50 W until the temperature reached at least 50°C. DNA samples, radiolabelled as described in following sections and suspended in Sanger dye (see Materials), were then loaded and run at 50 - 60°C, adjusting the power as required. At first gels were 'fixed' in 10% methanol, 10% HAc in water but this was found (by others) to be unnecessary and the practice was dropped. Gels were blotted and then dried at 80°C for usually 1 hr 15 min in a BioRad Model 583 gel drier attached to a vacuum system. Images were then obtained either by exposing the gel to X-ray film in a cassette and developing the film (with an X-OGRAPH Compact X2 automatic film processor) or by exposing the film in a PhosphorImager cassette (Molecular Dynamics) and scanning the (reusable) phosphor screen with a Molecular Dynamics Series 425S PhosphorImager™ and viewing the resultant computer files with ImageQuant™ software.

Large scale λ DNA preparation ('Maxiprep')

This 5-day method was used to produce large quantities of bacteriophage DNA for construction of new strains.

Day 1, Preparation: L broth was made up to the recipe given under Media above, but with the addition of 2·465 g $MgSO_4.6H_2O$ (final conc. 5 mM) and 20 ml 1 M Tris.HCl (pH 7·5) per litre, autoclaved and left in a 37°C room overnight along with 2l flasks (one per 250 ml lysate) and sterile 250 ml measuring cylinder. Plate lysates of the phage to be used were titred on the cells to be used for the lysate. Dialysis tubing was prepared as in Sambrook *et al.* (1989). CsCl solutions, pronase, DNase and RNase were made up as described above under Materials and overnight cultures of cells were set up allowing > 5 ml of culture per 250 ml of liquid lysate.

Day 2, Liquid lysis: the overnight cultures were diluted 50-fold in the warm L broth (5 ml of overnight culture in 250 ml of broth in 2l flasks to give plenty of room for

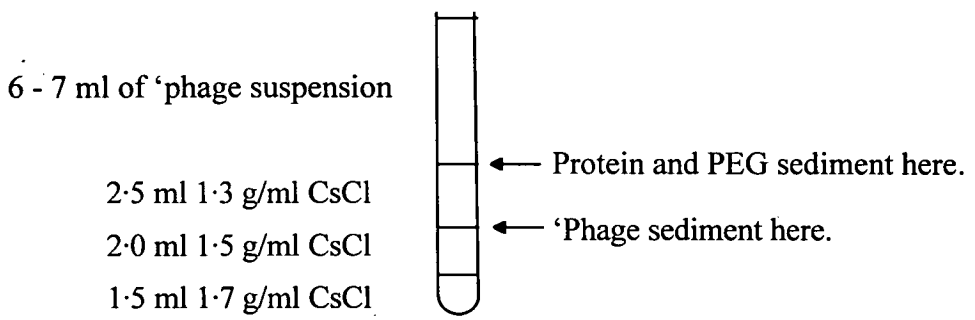
good aeration) and grown in orbital incubator at 37°C and OD_{650} monitored. An OD_{650} of 0·5 (about 2×10^8 cells/ml, which is reached at about 80 min) is the usual point of infection but for 'phage with long palindromes an earlier point is better such as 0·3 (about $1·2 \times 10^8$ cells/ml) which is reached at about 1 hr (see Allers, 1993). 'Phage were then added to an MOI (multiplicity of infection) of 0·1. The flasks were then returned to the orbital incubator at 37°C, shaking at about 200 r.p.m.. The cells continue to grow for 3 - 4 hr (reaching an OD_{650} of 2 - 3) before lysis starts. However, lysis is not complete until about 5 - 8 hr after infection, by which time the OD_{650} has fallen to 0·5 - 1.

When the OD_{650} dropped no further (or started to rise again) 500 µl chloroform were added and shaking continued for a further 10 min to lyse any remaining cells. The lysate was then decanted into 250 ml sterile Nalgene bottles, placed on ice for 10 min and centrifuged for 15 min at 10 krpm at 4°C (Sorvall® centrifuge, GSA rotor) to remove cell debris. The 250 ml supernatant lysates were poured into a 500 ml conical flask and 25 µl DNase (10 mg/ml) and 25 µl RNase (10 mg/ml) added (*i.e.* 1 µg/ml final concentrations) and returned to the orbital incubator and swirled gently at 37°C for 30 min. 14·6 g portions of NaCl were then added to the lysates (to make a concentration of 1 M) to dissociate further bacterial debris from the 'phage and the flasks swirled for another 10 - 15 min to dissolve the NaCl. The lysates were then poured into fresh 250 ml Nalgene bottles and left at 4°C overnight.

Day 3, PEG precipitation: The lysates were centrifuged again for 15 min at 10 krpm at 4°C and the supernatants decanted onto 25 g portions of good quality PEG 6000 (to give a concentration of 10% w/v) and stirred gently at room temperature to dissolve the PEG, then decanted into Nalgene centrifugation bottles and left on ice for 3 - 4 hr to precipitate the 'phage. The bottles were then centrifuged for 15 min at 10 krpm at 4°C (in a Du Pont Sorvall® refrigerated superspeed centrifuge), the supernatants discarded. and the bottles left upside-down to drain for 10 - 15 min.

The precipitate, spread in an annuloid, partly on the bottom but mainly on the side of the bottle, was resuspended in 5 ml of 'phage buffer at room temperature by taping the bottle at a slight angle to the horizontal and shaking gently on a rotary gel-shaker tray for 1 - 2 hr. The suspension was transferred to a glass bottle and the centrifuge bottle was rinsed with a further 1 ml of 'phage buffer which was added to the rest of the suspension. 7 ml chloroform was added to the suspension and the bottle shaken for 15 - 30 sec and then centrifuged in a 'bench-top' centrifuge (MSE Centaur-2) for 10 min at full speed (4·5 krpm) to bring down precipitated PEG, and the aqueous layer transferred to a clean bottle and kept at 4°C overnight.

Day 4 - CsCl separation: CsCl 'step gradients' were set up by successively underlaying 2·5 ml of 1·3 g/ml, with 2·0 ml of 1·5 g/ml CsCl and 1·5 ml 1·7 g/ml CsCl solutions in 13·2 ml Beckman Ultra-Clear™ $^{9}/_{16} \times 3\frac{1}{2}$ in tubes using a Pasteur pipette and a 1 ml pipette-filler. Then the 6 - 7 ml of 'phage suspension was gently laid on the top to within 2 mm of the rim.



The tubes were balanced to ± 10 mg and centrifuged for 1 hr at 35 krpm at 18°C (in a Du Pont Sorvall® OTD-Combi Ultracentrifuge with a TH641 titanium swing-out rotor). Two white bands could then be seen as shown above by viewing against a black background with visible light illumination from the top.

During centrifugation, lengths of dialysis tubing (stored in 2% w/v $NaHCO_3$/1 mM EDTA) were given two 300 ml rinses with 1 mM $MgSO_4$, to overwhelm EDTA to preserve 'phage integrity, and clamped at one end with labelled clamps. 'Phage bands (about 1 ml) were removed through the sides of the tubes with hypodermic needles and 1 ml syringes and dialysed for 1 - 2 hr in 2 l of 'phage buffer at 4°C.

During dialysis a second step gradient was prepared for each 'phage, the same as the first. The contents of the dialysis tubing were loaded onto these and, to minimize 'phage loss, each piece of dialysis tubing was rinsed inside with 1 ml of 'phage buffer which was added to the rest in the respective ultracentrifuge tube. Ultracentrifugation was carried out as before and the 'phage bands removed again. This time PEG/protein bands should be negligible. This time the dialysis tubing was rinsed with sterile distilled water and 1 l of 1 × TE buffer per 'phage preparation was used for dialysis. The reason for using TE this time is the impending DNA preparation. As before dialysis was at 4°C for 2 hr, or left running overnight.

Pronase dialysis buffer was prepared thus: 20 ml 1 M Tris, pH 7·5, 20 ml 5 M NaCl, 4 ml 0·25 M EDTA, 200 ml 10% Triton X100, made up to 1 l with sterile distilled water. This was left at 37°C overnight.

Day 5 - DNA preparation: Pronase, 20 mg/ml, was added to the 'phage suspension in the dialysis tubing to a final concentration of 1 mg/ml, *i.e.* 1/19 volume, the dialysis tubing reclamped and dialysis continued against the warm buffer (made up the previous night) for 2 hr at 37°C to digest the 'phage coat. Then the contents of the dialysis tubing were poured into a 1·5 ml Eppendorf tube and 500 µl of 1 × TE used to rinse the inside of the tubing and added to the rest of the suspension. The total of a little under 1·5 ml was divided between three Eppendorf tubes so that each contained a little under 500 µl of 'phage DNA suspension. These were then extracted with phenol, which would disrupt any remaining 'phage heads, and then with phenol-chloroform and with chloroform-isoamylalcohol, to purify DNA from denatured protein.

By this time the volume of aqueous suspension in each Eppendorf tube was around 450 µl, making it just possible to perform ethanol precipitation in the tubes. The resulting DNA pellets were then each resuspend in 60 µl and pooled in one of the three tubes per 'phage preparation and then the DNA was isopropanol-precipitated (and rinsed as usual with 70% ethanol). The dried pellets were

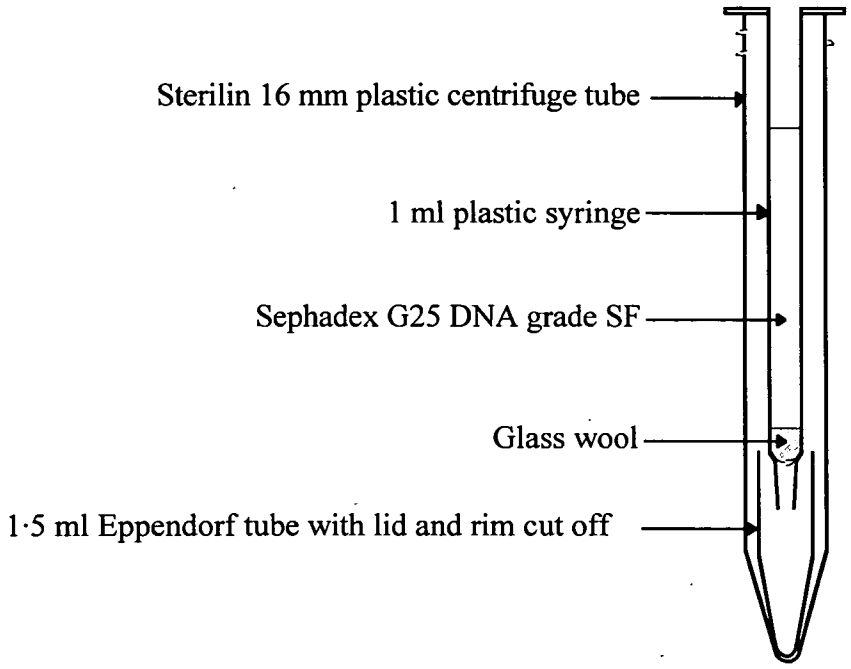
resuspended in 200 µl 1 × TE. DNA concentrations were then estimated by taking 10 µl of the suspension and adding it to 490 µl 1 × TE (*i.e.* 1/50 dilution), measuring the OD_{260} (Perkin-Elmer Lambda 15 spectrophotometer) and multiplying by 2,500 to get µg/ml (or ng/µl).

Checking oligonucleotide sequences by A+G Maxam & Gilbert sequencing

Normally one could rely upon commercially-produced oligonucleotides to have the sequence that had been ordered but this method was used to check some oligonucleotides when bacteriophage constructions behaved in a manner that suggested that they were not what had been intended. This made it possible to determine that oligonucleotides were the right lengths and had A and G nucleotides in all the right places, and that was sufficient. The method is derived from that of Maxam & Gilbert (1977). Purines are methylated with formic acid and the DNA chain is then cleaved on both sides of the modified bases by piperidine.

First, the oligonucleotide was 5' end-labelled (10 pmol of oligonucleotide DNA, 2 µl T4 DNA ligase buffer, 10 µCi of ^{32}P-γdATP and 2 µl T4 DNA polynucleotide kinase pre-diluted 1/25 with dilution buffer, *i.e.* about 2 units (see under Enzymes under Materials), made up to 20 µl with distilled water and incubated at 37°C for 45 min). The enzyme was then denatured by heating at 65°C for 20 min. The labelled oligonucleotide was then separated from unincorporated label by passing it through a 'spin column'. The latter was made by first packing a small wad of glass wool into the bottom of a 1 ml plastic syringe, then adding a slurry of Sephadex G25 DNA grade SF (the powder suspended in a little water by swirling in a small conical flask) with a Pasteur pipette. Water drips out of the bottom of the syringe as the slurry is being added. When the syringe was full to the brim it was centrifuged (at mark 3 on a Heraeus Christ GMBH bench-top centrifuge) for 5 - 6 min, suspended by its rim in a 16 mm Sterilin centrifuge tube and stabilized at its lower end by

insertion of the nozzle into a 1·5 ml Eppendorf tube with its lid and rim cut off as shown. Then, if the top of the packed Sephadex was below the 1 ml mark of the syringe, more Sephadex slurry could be added, but this is not usually necessary.



After removing water that had passed out of the column into the Eppendorf tube, the column was put back in place, the labelling reaction added carefully to the top of the column and the top covered with Parafilm. The arrangement was then centrifuged for 3 min at mark 3 on the Heraeus Christ bench-top centrifuge. (This particular centrifuge was used because it was the one designated for radioactive work. The setting and time were discovered by Lynne Powell by trial and error. The manual has long ago been lost and nobody here knows the speed.) The column was then removed and the eluate transferred to another Eppendorf tube for storage. 300 Ci of this labelled oligonucleotide was then taken for cleavage. This was measured out by putting 1 µl of the eluate into an Eppendorf tube and holding the bottom of the tube to a Geiger counter detector, adding the extra volume required and rechecking with the Geiger counter. The total was only a few microlitres. To this was added 2 µl calf thymus DNA (1 mg/ml, Sigma-Aldrich Company Ltd, Poole, Dorset), the

volume was made up to 10 µl with distilled water and the tube was put on ice for 5 min. Then 1·4 µl 100% formic acid was added and the reaction incubated at 37°C for 14 min. After cooling on ice for 10 min, 150µl of fresh 1 M piperidine was added, the tube sealed and held shut with a weight, and the reaction incubated at 90°C for 30 min and then cooled on ice again for 10 min.

The piperidine was removed by evaporation in a heated vacuum centrifuge (GeneVac SF50) for 2 hrs initially, then transferring to a fresh tube followed by resuspension with 100 µl water and re-evaporation three times for 30 - 60 min each time. Then the pellet was resuspended in 10 µl of stop solution. 2 µl of the suspension was run on a 12% polyacrylamide gel until the bromophenol blue front was not more than $10^{15}/_{16}$ inches (27·75 cm) from the wells (otherwise the one-nucleotide band, if there is one, runs off the gel), the gel dried and exposed to an X-ray film or PhosphorImager screen for about 3 days.

Preparation of bacteriophage packaging extracts

Most of the work described was done with stocks of these materials that had already been made up and stored at –70°C. They were made up as follows:

Buffer A

20 mM Tris, pH 7·5, 3 mM $MgCl_2$, 0.05% (v/v) β-mercaptoethanol, 1 mM EDTA.

Buffer M1

6 mM Tris, pH 7·5, 30 mM spermidine, 60 mM putrescine, 18 mM $MgCl_2$, 15 mM ATP, 0·2% v/v β-mercaptoethanol.

Sonicated Extract

An overnight culture of the lysogenic *E. coli* strain BHB2690 was made in 15 ml of L broth incubated at 30°C. Next day 10 ml of this was added to 500 ml L broth

and incubated, shaking at 30°C to an OD_{650} of 0·3. Then lysogenization was induced by transferring the culture to a shaking waterbath at 45°C for 15 min followed by vigorous shaking for 1 hr. The suspension was then cooled on ice for 10 min, split between centrifuge tubes, and centrifuged at 6 krpm for 6 min at 4°C (Du Pont Sorvall® refrigerated superspeed centrifuge, GSA rotor). The supernatant was removed and the cell pellets resuspended in 0·5 ml volumes of buffer A, transferred to a single 30 ml Nalgene polypropylene bottle, diluted with a further 2·6 ml of buffer A, and the suspension sonicated in 5 - 6 ~3-second bursts with 30 sec rests in between until the suspension was no longer viscous. It was then centrifuged again at 6 krpm for 6 min at 4°C (Du Pont Sorvall® refrigerated superspeed centrifuge, SS34 rotor) and 50 µl volumes of the supernatant transferred to precooled Eppendorf tubes, frozen with liquid nitrigen, and stored at –70°C.

Freeze-Thaw Lysate

An overnight culture of the lysogenic *E. coli* strain BHB2688 was made in 40 ml of L broth incubated at 30°C. Next day 10 ml of this was added to each of three volumes of 500 ml L broth and incubated, shaking at 30°C to an OD_{650} of 0·3. Then lysogenization was induced by transferring the culture to a shaking water-bath at 45°C for 15 min followed by vigorous shaking for 1 hr. The suspension was then split between centrifuge tubes, and centrifuged at 10 krpm for 10 min at 4°C (Sorvall® centrifuge, GSA rotor). The supernatant was drained and the cell pellets resuspended in 0·5 ml volumes of cold 10% sucrose, 50 mM tris, pH 7·5 and pooled in one Oak Ridge ultracentrifuge tube. To this was added 150 µl of freshly prepared lysozyme (2 mg/ml in 0·25 M tris, pH 7·5) and the suspension mixed gently and then frozen on liquid nitrogen, then thawed, first at room temperature then at 4°C until completely thawed, then cooled on ice. 150 µl of buffer M1 was added and the suspension mixed gently and then centrifuged at 40 krpm for 1 hr at 4°C (Du Pont Sorvall® OTD-Combi Ultracentrifuge, Ti50 rotor). 55 µl volumes of the supernatant

were transferred to precooled Eppendorf tubes, frozen with liquid nitrogen, and stored at −70°C.

Cloning oligonucleotide inserts in the palindrome centre of bacteriophage λ DRL167

Trinucleotide repeat and other sequences to be tested for their hairpin folding ability were inserted into the unique *Sac*I site at the centre of the 462 bp perfect palindrome of λDRL167. Double-stranded DNA inserts containing the test sequences were made by annealing complementary oligonucleotides. They were designed so as to destroy the *Sac*I site and provide a new restriction site, *Bsa*I, for identification of successful ligation products. The sequences of the inserts are listed in Table 2.1. The oligonucleotides were not phosphorylated so as to avoid multiple insertion.

All of the oligonucleotides used for inserts directly into DRL167 were manufactured by Oswel and came suspended in sterile distilled water at concentrations of 18 - 70 μM (18 - 70 pmol/μl). As they were palindromic, care had to be taken to promote the formation of duplex DNA rather than hairpins formed from single strands. This involved (i) annealing at high concentration and diluting later, and (ii) slow cooling. Equimolar proportions of the two oligonucleotides to be annealed were mixed in an Eppendorf tube to give a total volume of 64 μl. (For each oligonucleotide the volume to use is $\frac{C_{other}}{C_1 + C_2} \times 64$ μl, where C_1 and C_2 are the concentrations of the two oligonucleotides and C_{other} is the concentration of the one not being pipetted at the time, *i.e.* for oligonucleotide 1, C_{other} is C_2.) To this were added 8 μl of 10 × TE and 8 μl of 100 mM NaCl to give a total volume of 80 μl 1 × TE, 10 mM NaCl. The tube was then put into a polystyrene floater and placed on the surface of boiling water in an almost full 1 l beaker. The heat was then turned off and the beaker quickly covered with aluminium foil and then left to cool slowly to

room temperature overnight. In the morning the beaker was put in the 4°C room to cool further.

In the meantime, λDRL167 DNA was digested with *Sac*I. Usually several new 'phage were constructed at the same time, so enough DNA was digested for all of them. The volume of DNA used depended upon the concentration of the 'phage DNA maxiprep and the number of 'phage being constructed but was usually about 6 - 12 µl so could be digested in a 10 - 20 µl final volume. It was not necessary to destroy the enzyme before the next step - ligation with insert DNA - because the insert destroyed the *Sac*I recognition site. Then 10 µl of each annealed oligonucleotide suspension was diluted 1 in 100 with ice-cold 1 × TE, 10 mM NaCl and individual ligations were set up with an approximately 5-fold molar excess of insert over 'phage genome DNA. This worked out conveniently to the use of 1 µl of the diluted insert DNA with 5 µl of 'phage DNA digest. After adding the 'phage and insert DNA, ligase buffer and water, the tubes were incubated at 37°C for 2 min to melt cohesive ends and then returned immediately to ice before adding the ligase and incubating for ligation. Since half the volume of a 10 µl ligation reaction was contributed by the 5 µl of 'phage DNA digest, this contributed 0·5 × *Sac*I buffer. To cope with this, two strategies were tried. One was to ignore this fact and still add 1 µl of 10 × ligase buffer. This meant that enough ATP was added, in the ligase buffer, but the salt concentration was a bit higher than recommended for the ligase. The other was to add only 0·5 µl of 10 × ligase buffer and to add a little extra ATP solution (1 µl of 5 mM ATP). Since both worked very well, no attempt was made to find an optimum. Usually the ligations were incubated overnight.

The ligase was then denatured by placing the tubes in a heating block at 70°C. The block was then switched off and left to cool slowly till close to room temperature to allow reannealing at the centre of the palindrome as the oligonucleotides were only ligated at their 3' ends because they lacked 5' phosphates. Then the whole block was placed on ice. The products of this process were then

redigested with *Sac*I to cleave any 'phage DNA that had been ligated without an insert so that it would not be packaged. For this reaction, as for the ligation, the DNA was not ethanol precipitated and resuspended in the new buffer. Instead, 12 µl of 1 × *Sac*I buffer was added to the ligation reaction and 5 units of *Sac*I. The particular *Sac*I buffer used was that of NBL Gene Sciences Ltd and the ligase buffer was from New England Biolabs. Both have 10 mM $MgCl_2$ but the former has 33 mM Tris acetate, 66 mM KAc while the latter has 50 mM Tris.HCl.

In vitro packaging of the λDNA construct

After the digestion the tubes were again left in the heating block to cool slowly to room temperature and then the whole dry block was put on ice. Then the DNA was ethanol-precipitated, resuspended in 5 µl of 1 × TE and packaged *in vitro*. To the 5 µl of DNA suspension were added 7 µl of *in vitro* packaging buffer A, 2 µl of buffer M1, 10 µl of sonicated extract and 10 µl of freeze-thaw lysate. The mixture was incubated at room temperature (usually about 26°C) for 1 hr then diluted to 500 µl with 'phage buffer and put on ice. Then packaged 'phage were plated. Usually 1 µl of each diluted packaging reaction was sufficient to give adequate numbers of plaques, not too closely spaced. (JC9387 plating cells and BBL plates were used.)

'Phage selection and purification

From these plates, usually 26 plaques were picked for each new 'phage construct, using sterile toothpicks, and left in 1 ml volumes of 'phage buffer for at least 2 hr at 4°C for diffusion of the 'phage into suspension. In the meantime, plating cells were made from overnight cultures of *E. coli* strains R594 and N2364 (plating cells of JC9387 had already been made the day before) and grids of squares drawn on the bottoms of three BBL plates for every two 'phage constructs and the squares numbered and labelled. The grids had 52 squares so could accommodate spots of two

sets of 'phage suspensions. Lawns of JC9387, R594 and N2364 were then poured in the usual way - 2·5 ml of BBL top agar added to 0·25 ml of plating cells - but without any 'phage added to the cells. After allowing 5 min for the agar to set, a 2 µl drop of each 'phage suspension was placed on its respective grid square on each of the three bacterial host strains. Usually the same (yellow) pipette tip was used for each of the three drops and they were placed in the order R594 (*rec*$^+$), N2364 (*sbcC*), JC9387 (*recBC, sbcC*) to go from the least permissive to the most permissive. The plates were then incubated overnight.

Next day the plates were examined for 'phage suspensions that grew well on JC9387 but did not grow on R594 (any 'phage growing on R594 must have lost its palindrome) and usually three of these were chosen for each construct (the growth on N2364 sometimes helped the decision). These were then plaque purified by plating the suspensions corresponding to the chosen spots on BBL plates, picking one plaque from each and making suspensions from them and plating on BBL plates again. From these plates in turn, one plaque was picked and used for plate lysis.

Plate lysis of palindrome-bearing 'phage

Fresh JC9387 plating cells were prepared. In the meantime, one plaque was picked from the latest plaque-purification plates and left to diffuse in 1 ml of 'phage buffer for 1 hr at room temperature. L agarose top and bottom agar, made up the day before and kept overnight at 60°C, were cooled to 46°C and the additions (see under Media under Materials) were made to the bottom agar. The bottle was swirled to mix the contents and the plates were poured immediately and spread out to cool as they were to be used within one hour. 200 µl of undiluted plaque suspension was plated with JC9387 cells and L top agarose and the plates were incubated at 37°C until covered with tiny plaques just touching one-another. This was usually after about 6 - 7 hr incubation. 4 ml of 'phage buffer was then laid onto each plate and ideally left (at room temperature) for 30 min or more to soften the top agarose. Then the top

agarose was mashed with the side of the end of a 10 ml glass pipette and then sucked up and transferred to a McCartney Bottle. A further 2 ml of 'phage buffer was used to rinse the surface of the bottom agar and added to the rest. After harvesting all the plates, 50 µl of chloroform was added to each bottle and it was shaken up and kept at 4°C overnight for 'phage to continue diffusing out of the agarose. Next day the bottles were centrifuged at full speed in a 'bench-top' centrifuge (about 4·5 krpm) and the supernatants were transferred to ¼ oz bijoux bottles and kept at 4°C. Portions of these lysates were used for DNA minipreparations, to check the 'phage constructions, and the remainders were kept as 'phage stocks.

'Phage strain nomenclature

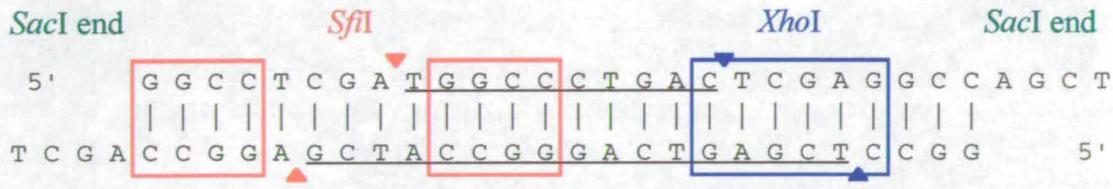
The original working name of each 'phage constructed in this work was of the form x,y where x is the construct number and y the plaque number in the order of plaques picked after plating the newly-packaged 'phage. For the very first constructions, oligonucleotides were just ordered to make inserts containing one d(CAG)·d(CTG) or one d(GAC)·d(GTC) trinucleotide in the centre. 'Phage containing these inserts were numbered respectively 1 and 2, but for these first constructions only, two controls were run, both using the d(CAG)·d(CTG) insert, one with no ligase and one in which the DNA was not recut after ligation and these were numbered 3 and 4 respectively. After plating the packaged 'phage, only a few plaques from 3 and 4 were picked and replated for selection but one of those, 4,8 was positive and was used for plaque size quantification, hence there being two different numbers for d(CAG)·d(CTG) 'phage in Table 2.1. Thereafter, constructs were numbered consecutively in the order that they were made. Later, 'phage that had been shown to be what they were supposed to be were given laboratory numbers, headed by the letters DRL and these were not necessarily allotted in the same order. Thus the first ten 'phage were made in the order $d[(CAG)\cdot(CTG)]_1$, $d[(GAC)\cdot(GTC)]_1$, $d[(CAG)\cdot(CTG)]_2$, $d[(GAC)\cdot(GTC)]_2$ .... $d[(CAG)\cdot(CTG)]_5$,

d[(GAC)·(GTC)]$_5$, but were later given laboratory numbers in the order d[(CAG)·(CTG)]$_{1-5}$ = DRL220 - DRL224 and d[(GAC)·(GTC)]$_{1-5}$ = DRL225 - DRL229.

Cloning oligonucleotide inserts in the palindrome centres of bacteriophage λ DRL257 and DRL258

The general design

DRL257 and DRL258 were made by ligating the following insert into the *Sac*I site of DRL167 by the above described procedure:

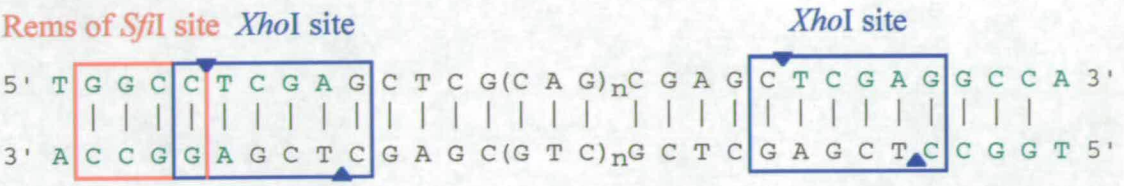


In common with the other inserts used with DRL167, it destroyed the *Sac*I site. The insert could go into DRL167 in either orientation and isolates of the resulting 'phage with each orientation were identified as described in Chapter 7. The isolate with the above orientation, with the *Sfi*I site nearer to the left end of the bacteriophage genome (orientation A), was named DRL257 and isolates with the opposite orientation (B), were named DRL258. Double digestion of either 'phage genome with *Sfi*I and *Xho*I would excise the underlined fragment. As this left overhanging ends in opposite senses, inserts into either of these 'phage could go in only one orientation, determined by the parent 'phage.

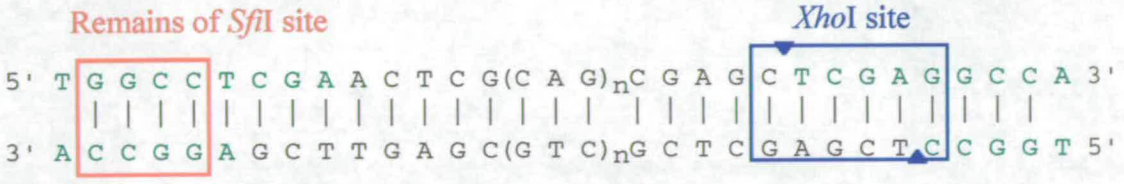
The inserts had one long oligonucleotide and one short one. Table 2.2 lists the inserts of those 'phage which have been shown to have the intended structure. (Other 'phage that were constructed but have not been tested, or for which no correct isolate has yet been identified, are mentioned in Chapter 7.) The inserts were of two types, 'symmetrical' and 'asymmetrical', illustrated below as inserted into the DRL257 palindrome (green). Both types of insert destroy the *Sfi*I recognition site as

they lack the required second GGCC motif. The 'symmetrical' inserts restore perfect symmetry to the palindrome outside the central trinucleotide test sequence but at the cost of creating an *Xho*I recognition site on both sides. The 'asymmetrical' inserts make a single base-pair difference between the two sides so that an *Xho*I site is only restored on one side.

Symmetrical insert



Asymmetrical insert



$(CAG)_n \cdot (CTG)_n$ represents the test trinucleotide repeats. $(GAA)_n \cdot (TTC)_n$ repeats were also used in asymmetrical inserts.

Cloning procedure

This was the same as for making new constructs from DRL167 except for the following modifications: Oligonucleotides for some of these inserts were ordered from Perkin-Elmer (Warrington). They came suspended in 20% (v/v) acetonitrile/water at only about $^1/_5$ of the concentration of Oswel oligonucleotides or less; they ranged from 2·0 - 12·1 pmol/µl. Therefore, after annealing, dilution by 1 in 20 would give roughly the same concentration as annealed Oswel oligonucleotides diluted 1 in 100. One feature of *Sfi*I/*Xho*I digests of DRL257 and DRL258 DNA not present in *Sac*I digests of DRL176 is the small fragment underlined above. This will compete with the new insert DNA if not removed. Initially a commercial kit was

tried for this separation, but when it was found that some more time was going to be required to get a good yield of 'phage 'arms' (halves of the genome), it was decided instead to ethanol-precipitate the digested DNA and then use a 100-fold excess of insert over 'phage arms instead of the 5-fold excess used before. This meant using twenty times as much insert as before, *i.e.* 1 µl of undiluted annealed Perkin-Elmer oligonucleotides. Later on Genosys (Cambridge) oligonucleotides were used. These came lyophilized and even when resuspended in 1 ml, their concentrations were still about five times higher than those of the Oswel oligonucleotides; the range was 124 - 288 pmol/µl. They were accordingly resuspended in 1 ml (of 1 × TE and left for 1 hr at room temperature) and, after annealing in the usual way, the DNA was diluted 1 in 20 (with the usual 1 × TE, 10 mM NaCl) and 1 µl taken for ligation.

Digestion of the parent 'phage (DRL257 or DRL258) DNA could not be done in a simultaneous double digestion because the optimum incubation temperature of *Sfi*I is 50°C whereas for *Xho*I it is 37°C. The *Xho*I had been obtained from New England Biolabs and the *Sfi*I was obtained from NBL Gene Sciences Ltd. The buffers recommended by the respective companies for their enzymes had the same concentrations of NaCl, Tris, and $MgCl_2$ but the one for *Xho*I had pH 7·9 and the one for *Sfi*I pH 8·4. However, New England Biolabs also produce *Sfi*I and they recommend the same buffer for both enzymes (NEB Buffer 2, pH 7·9) so that was used. Digests were set up with *Xho*I only and incubated at 37°C for at least 2 hr and then the condensation was spun down, *Sfi*I was added and incubation continued at 50°C for at least another 2 hr, usually spinning down condensation about half way through to prevent the reactions from becoming too concentrated.

Since the two arms of the parent 'phage have different 'sticky ends' and the little central piece from the parent 'phage (if still present after ethanol precipitation) was vastly outnumbered by the new insert DNA, it should not have been necessary to recut after ligation. However, neither *Sfi*I nor *Xho*I is a very efficient enzyme so there were bound to be some 'phage arms with an attached central DNA piece and these could ligate to the other 'phage arm in a first order reaction whereas

construction of the intended new 'phage genome required the ligation of three molecules. Since the *Xho*I site was not disrupted by the inserts, redigestion was carried out with *Sfi*I only (with its own manufacturer's recommended buffer, nbl Buffer 12 with pH 8·4).

'Phage DNA miniprep

After separating plate supernatant lysates from agarose, as above, 1 ml of each was immediately taken into Eppendorf tubes for DNA minipreparation. They were first digested with 5 µl DNase I and 5 µl RNase A (prepared as above) at 37°C for 30 min, then centrifuged for 1 min in a microcentrifuge and the supernatants added to 500 µl volumes of 30% PEG 6000, 3M NaCl in Eppendorf tubes on ice, inverted several times to mix, and left on ice for 2 - 4 hr. They were then centrifuged for 15 min at 4°C. The supernatants was removed and the tubes briefly centrifuged again and remaining liquid removed and then the pellets were resuspended gently in 200 µl volumes of 1 × TM. 200 µl of chloroform was then added to disrupt the 'phage coats and the tubes were vortexed briefly and centrifuged for 5 min at room temperature. After removal of the aqueous layers to new tubes, phenol, phenol-chloroform and chloroform-isoamylalcohol extractions were successively carried out as already described, followed by ethanol precipitation and then by isopropanol precipitation. The final dried precipitates were resuspended in 20 µl 1 × TE. This provided enough DNA for the following tests of 'phage construction.

Testing 'phage DNA for presence of inserts by restriction digestion

4 µl of miniprep DNA was taken for each of two restriction digests. For 'phage constructed directly from DRL167 DNA these were of *Sac*I, whose recognition site is abolished if an insert is present, and *Bsa*I, for which a new site is introduced by the insert (actually two sites but so close together that it makes no

difference). The *Sac*I digests were carried out in 10 µl (at the optimum 37°C) but at the optimum temperature for *Bsa*I, 55°C, the tube contents completely evaporated from the bottom of the tube with this volume. Mineral oil could have been used but instead it was found satisfactory to use a reaction volume of 20 µl and to spin down condensation once in the middle of a 2 hr incubation. The digests were then examined by agarose gel electrophoresis (*q.v.* above) with 8-well minigels or 14-well 'midigels' as numbers dictated. 0·5% agarose was used for the midigels with 100 ml volume for the 10 µl digests (made up to 12·5 µl with loading buffer) and 125 ml volume for the 20 µl digests (made up to 25 µl with loading buffer). If minigels were used then the agarose concentration used was 0·6% (for strength) and the volume was 25 ml for the 10 ml digests and 35 ml, with a thick-toothed comb, for the 20 ml digests. Digestion with *Sac*I produced two fragments if there was no insert but no cleavage if there was an insert. Cleavage with *Bsa*I produced three fragments if there was no insert and four if there was one. (Examples are shown in Chapter 4.)

For constructs from DRL257 and DRL258, the digests were with *Sfi*I and *Xho*I. As the optimum temperature for *Sfi*I is 50°C, digests with this enzyme were carried out in 20 µl volumes as for those of *Bsa*I while *Xho*I digests were carried out at 37°C in 10 µl and gels were run exactly as for the other enzymes. Cleavage with either *Sfi*I or *Xho*I gave a result exactly the same as that for cleavage with *Sac*I, *i.e.* cleavage in the centre of the palindrome and nowhere else in the genome. *Xho*I should cleave whether or not there was an insert and absence of cleavage would suggest that the recognition site had been lost by a deletion of the centre of the palindrome. Provided that cleavage with *Xho*I did occur, absence of cleavage with *Sfi*I indicated that there was an insert. Several digests with the same enzyme were always done at once so there was adequate control for failure of the enzyme.

Checking the sizes of inserts by PAGE

Having established that insertion had occurred, it was still necessary to verify that the insert was of the intended size. It was not possible to sequence across the 462+ bp palindrome because of secondary structure, and with constructs made directly from DRL167 and ones made with symmetrical inserts in DRL257 and DRL258 it was not possible to cleave the DNA to one side of the insert only and so to sequence only one side of the palindrome with the insert. It was however possible to compare the lengths of the palindrome without and with inserts of different sizes.

The palindrome has *Eco*RI sites at its ends. 10 µl of miniprep DNA samples of the 'phage under test were digested with *Eco*RI in total volumes of 20 µl. The fragments were then 3'-end-labelled as follows. (0·5N + 0·5) µl of [α^{35}S]dATP (10 Ci/µl), (where N = the number of samples) was dispensed into an Eppendorf tube and to this was added the same volume of unlabelled dTTP. 1 µl of this mixture was then added to each 20 µl *Eco*RI digest followed by 1 unit of Klenow enzyme (which originally came at 1 U/µl but later at 2 U/µl) and incubated for 15 min at room temperature. Then 0·5 µl of 'chase' (12·5 mM of each of the four dNTPs) was added to each reaction followed by another 1 U of Klenow enzyme and incubated for a further 10 min at room temperature. Then 75 µl of ice-cold 1 × TE was added, followed by extraction with 100 µl phenol-chloroform and then with 100 µl chloroform-isoamylalcohol. The DNA was then ethanol-precipitated and resuspended in 10 µl of Sanger dye.

2 µl of this suspension was then subjected to polyacrylamide gel electrophoresis for 6 - 7·75 hr at 50 - 60°C and the result viewed by autoradiography or phosphorimaging.

Checking for presence and sequence of inserts in λ DRL257 and DRL258 by automated sequencing

'Phage constructs from these parents with asymmetrical inserts had a single *Xho*I site to one side of the central test sequence (see above under cloning in these 'phage). Cleavage with *Xho*I therefore divided the 'phage genome into two parts each having one half of the palindrome, which, separated from its other half, did not cause a problem in sequencing, and one of these halves had the inserted test sequence at the end. After digestion with *Xho*I, the DNA was ligated to the following small piece of DNA constructed by annealing oligonucleotides:

```
5' T C G A C A G A C T G A T G A C G A T T A C C C 3'
           | | | | | | | | | | | | | | | | | | | |
     3' G T C T G A C T A C T G C T A A T G G G 5'
```

This has an *Xho*I 'sticky end' and the 5' base of the longer oligonucleotide was phosphated so both strands ligated. The piece of course ligated onto both arms of the palindrome but the sequence of a single arm could be amplified by PCR between a primer matching this piece and a primer recognizing a unique sequence outside the required arm of the palindrome. The three primers were:

Ligation piece:	5'GGGTAATCGT CATCAGTCTG TCG	(named *Xho*LigPri)
Left arm	5'AACCGAAGAA TGCGACACTG	(named PalJDleft)
Right arm	5'GAACAACCTG ACCCAGCAAA	(named PalJDright)

The left primer corresponds to bases 21,054 - 21,073 of the λ genome and gives a product of 432 bp with no insert or (346 + 3N) bp with an insert with N central trinucleotides. The right primer is complementary to bases 26,205 - 26,186 of the λ genome and gives a product of 356 bp with no insert or (370 + 3N) bp with an insert.

The oligonucleotides used to make the ligation piece were both from Perkin-Elmer but the longer (phosphated) one was supplied at an unusually high concentration for Perkin-Elmer (54 μM). They were annealed in the usual way and

diluted 1 in 100, giving a theoretical concentration of the double stranded DNA of 75·6 fmol/µl. 4 µl of miniprep DNA was digested with *Xho*I in a 10 µl volume and half of this was ligated to 1 µl of the diluted ligation-piece DNA in a volume of 10 µl. The ratio of the ligation piece to the genomic fragments was always unknown because the concentration of the miniprep DNA was unknown but as long as a few molecules ligated there would be a template for PCR so it was not important. (Since PCR was always between the ligation piece and a single primer for the genomic DNA there was no risk of a spurious product from re-ligated or uncleaved genomic DNA.) 1 µl of the ligation reaction was used as template in a 50 µl PCR reaction with 25 pmol of each primer, 10 nmol of each dNTP and 1 U *Taq* polymerase (Boehringer-Mannheim) with the supplier's reaction buffer. The program was 5 min initial melt at 94°C (then holding at this temperature whilst the polymerase was added) followed by 30 cycles of 30 sec at 94°C, 15 sec annealing at 55°C, and 40 sec extension at 72°C.

The PCR products were then purified using a QIAquick PCR Purification Kit (QIAGEN®). The double-stranded DNA is bound to a silica-gel membrane in a spin column, through which primers and other constituents pass, then rinsed with an ethanol-containing buffer and eluted with 10 mM Tris·Cl, pH 8·5. 30 µl of this buffer was used for the elution and it was spun through the column twice for a better yield. 10 µl of the resultant DNA suspension was then used for concentration measurement by spectrophotometry.

The DNA was then sequenced using the genomic PCR primer. (Since the insert sequence was at the other end of the PCR product immediately adjacent to the ligation piece DNA, the other primer could not be used for sequencing.) Sequencing was done with the ABI PRISM™ dRhodamine Terminator Cycle Sequencing Ready Reaction Kit with AmpliTaq® DNA Polymerase, FS. The enzyme is a variant of *Thermus aquaticus* DNA polymerase that contains a point mutation at the active site. This is said to result in reduced discrimination against dideoxynucleotides, leading to a more even peak intensity pattern. A second mutation virtually eliminates

5'→3' nuclease activity. The procedure is very simple. A single tube contains dNTPs, including dichlororhodamine-labelled dideoxy terminators of four different colours, polymerase and reaction buffer. To 8 µl of this are added template DNA, 3·2 pmol of primer and water to a total volume of 20 µl. The amount of template DNA recommended in the instruction manual is 30 - 90 ng if PCR product is used. Our automatic sequencer operator recommended 80 ng/kb. My experience was that at least twice as much as this is required to get a good signal to noise ratio. The reaction mixture is overlain with 40 µl of mineral oil and subjected to 25 cycles of 96°C for 30 sec, 50°C for 15 sec, 60°C for 4 min in a PCR machine.

The kit instructions are then to pipette the entire 20 µl reaction volume from under the 40 µl of oil with as little oil as possible and to add it to 2 µl of 3 M NaAc (pH 4·6) and 50 µl of ethanol in another tube. This is absurd. The oil has an affinity for the plastic (yellow) pipette tip and the little bubble of aqueous reaction nearly always skips out of the way of the end of the tip and slides up the side of the tube (try it). If one does manage to pierce the aqueous droplet at the bottom of the tube, a cone of oil surrounds the pipette tip and follows it down. At the very best, one can only hope to get about 16 µl of the aqueous phase with about 5 µl of oil. I therefore adopted a different strategy. I removed as much oil as possible first, discarded the tip, and then added 30 µl of water. It was then possible to remove all or nearly all of the 50 µl aqueous phase from under the thin layer of oil with a fresh pipette tip with very little oil taken up, as with a normal PCR reaction, and to this could then be added 5 µl of 3 M NaAc and 125 µl of ethanol. The rinsed and dried ethanol precipitate was then handed to our excellent automatic sequencer technician, Nicola Preston, who did the rest, *i.e.* resuspended it and subjected it to PAGE in a single lane on a Perkin-Elmer ABI Prism™ 377 DNA sequencer which reads the peaks of fluorescent light from the terminators as they pass a fixed level in the gel. The machine produces two computer files for each lane, one containing the interpreted sequence as text and the other showing all the dye peaks with interpreted sequence

above the peaks. It is always worthwhile to inspect the latter carefully, even when the text file contains no 'N' representing an undecided base. Peaks of 'noise' may often be distinguished from underlying true sequence peaks by their shape.

Chapter 3

Methods II. Plaque size quantification (PSQ)

The early protocol

This is the protocol almost as I inherited it from my predecessors, Angus Davison and Thorsten Allers, and was used to obtain the results described in Chapter 4. I made a few slight modifications which are mentioned.

Plates and plating

PSQ top and bottom agar was made up as in Chapter 2 (Materials section). The quantities given allow for 12 plates per bottle of bottom agar. Time and space, particularly when pouring the top agar, dictated that 48 was about the maximum comfortable number of plates to be *used* for one plaque assay. The number of plates poured had to provide for the removal of four plates per stack because of their rates of drying (see below) and for some spares to cover contamination of a few plates during drying, and for occasional accidents. 90 mm plastic Petri dishes from 'philip harris Scientific' were used. These come in bags of 15 and nearly always the bottom plate in each bag has concentric scratches on the bottom. It is necessary to remove any such plates before pouring because plaque size is measured with illumination from beneath the plates. The plates were poured to a volume of exactly 40 ml using a sterile 25 ml pipette. They were poured one on top of another in stacks of > 20 plates and left to dry for three days.

In the meantime, plate lysates of the 'phage strains to be investigated were plated (by serial dilution down to 10^{-7} and plating 10 µl) and 5 plaques were picked from each plate with sterile Pasteur pipettes, 'phage suspensions were made from each of them and these were diluted 1 in 100 and 10 µl of this dilution plated on

JC9387 cells and the resulting plaques counted to calculate the volume of each dilute suspension required to give 350 plaques/plate for the plaque size assay. About 60 - 100 plaques on each plate were to have their areas measured using an image analyser (see below). If the plaque density was too high, many pairs of plaques would overlap and so be unmeasurable. If the plaque density was too low, there would not be very many plaques in one field of vision (by the electronic camera) and multiple fields would have to be measured, taking more time, or the magnification would have to be altered so that the plaques appeared small on the screen, and then the measurement was less accurate. The optimum density would vary with the size of the plaques and the original protocol I was given was to plate about 300 p.f.u./plate for strains producing small plaques and about 150 -200 for strains producing large plaques, but these densities were rather low. 350 plaques/plate was sufficient for all sizes without any requirement to know in advance how large the plaques would be on N2364 cells on PSQ agar.

Then, when the PSQ plates were 3 days old, the top two and (when reached) the bottom two plates of each stack were set aside because of being more or less dry than the rest, thereby affecting plaque size. The rest were dealt, from the top, like playing cards into the required number of piles, one pile for each 'phage isolate being assayed at the time. (Any contaminated plate encountered was taken out and dealing went on.) When the last-but-two plate of the first stack had been dealt, dealing was continued from the next stack of plates in the same way until each new pile had 5 plates. Previously (Davison & Leach, 1994b) the plates were randomly allocated to different piles. This may have been satisfactory when only four strains were being compared, with twelve plates per strain, but I decided from the start that if there was enough difference in the rate of drying between different plates in a stack to require that the top two and bottom two should not be used then randomization was probably not good when only five plates were to be used per strain. A bias could be introduced by one strain happening to have several plates from near one end of a stack and another having several from near the other end. Dealing the plates in

rotation might itself introduce a bias if there was a gradient of dryness from top to bottom of the stack but if so that could be detected and corrected while random allocation of plates could introduce untraceable noise.

The plates were then labelled on the side walls because labelling underneath would obscure plaques when they developed. Then 250 µl volumes of N2364 plating cells, made earlier that day and kept at 4°C in the meantime, were inoculated with the calculated volumes of the 'dilute 'phage suspensions, and after 'phage adsorption, poured onto their respectively-labelled plates with 2·5 ml of PSQ top agar and incubated overnight at 37°C.

Plaque measurement

Plaque areas were measured initially with a Quantimet 970 digital image analyser (Cambridge Instruments). Illumination was provided by a light box some distance under a glass stage on which the plate was placed with its lid off, and an image of an area of the plate was presented to the analyzer via a Chalcinon video camera. The intensity of illumination was adjusted automatically by the image analyzer so that the peak brightness of the plaque images was equal to a video signal of 1 volt. The height of the stage above the light box and of the camera above the stage (and the focus of the camera) were adjusted to get a good clear image with reasonably sharp edges to the plaque images and a field area that struck a balance between showing a lot of plaques but rather small (and therefore more subject to measurement error) and showing large plaque images but not very many. At first three fields were measured per plate and therefore one wanted to have at least 20 good plaques per field. Once all the settings had been decided upon, the system was calibrated by placing a scale in millimetres on the stage on top of a Petri dish lid which gave it just about the same height above the top of the stage as the surface of the agar in a dish. Then, using the 'puck' (a sort of computer mouse), a line was drawn on the screen between two points on the image of the scale and the distance that this represented was typed into the computer. At the smallest field size used,

the screen corresponded to a rectangle of 2 × 1·6 mm on the plate. Then all distances were kept the same until all plates in a batch had been measured. (Inevitably by the next time one used the analyzer several weeks later, everything had been altered and the process had to be gone through again.)

The screen showed two superimposed images of the plate. The underlying one was a 'grey-scale' image which showed the plaques, in 'black-and-white' more or less as they appear to the eye, and the other was the digital image to be processed. The processing of the image was carried out using a little program (known as a Quips routine) written by Dr. C.E. Jeffree (Science Faculty Electron Microscope Facility, University of Edinburgh) and called "PLAQUE".

The digital image consisted of 896 × 704 pixels with 256 'grey levels' per pixel. It was actually an all-or-nothing image. In any position on the screen a single-intensity (bright yellow) pixel would appear if the light coming through the plate at that point was above a certain threshold intensity and no pixel would appear if not. The digital image therefore consisted of bright patches where light was shining through the plaques and from this the computer would make the area measurements. However, the bacterial lawn around a plaque is not, it seems, of even thickness up to a sudden cliff at the edge of the plaque, and the floor of the plaque is not of even translucency. Both grade away from the edge of the plaque. The first task therefore was to adjust the threshold so that the digital image corresponded as closely as possible to the perceived edge of the plaques. This was easiest if the original settings had resulted in the background (bacterial lawn) appearing very dark on the grey-scale image. In that case the edges of the plaques appeared quite sharp. If the background was lighter the plaques could be seen to have a (fairly steep) graduated edge and it was necessary to decide exactly where on this slope one was going to set the edge of the digital image and then be very careful to be as consistent as possible throughout all measurements with other fields.

At this point with the first field in view it was best to make final adjustment of the position of the light box. Imperfect positioning could lead to quite uneven

distribution of illumination over the field. This could result in the digital images of plaques at one side or corner of the screen being much smaller than the grey plaque images while at the opposite side or corner of the screen extraneous pixels appeared all over the background, where light shone through any inconsistencies in the bacterial lawn, and the digital images of the plaques overflowed the grey images and even merged together with those of other plaques. Moving the light box could vastly improve this but never completely eliminated it. One therefore strove to obtain as little difference as possible across the screen and adjusted the threshold using plaques in the centre of the field, giving a balance between plaques just a little bit small on one side and a little bit too big on the other.

The next operation was the removal of single pixel noise (spots on the background) by an erode procedure followed by a dilate procedure, *i.e.* pixels were removed from the edges of all objects in the digital image and then they were put back onto all objects that still existed and so dots of light that were not plaques disappeared. The same mechanism, with a variable number of erosions followed by an equal number of dilatations could be used to separate plaques that were just touching because if an object broke into two when eroded the two resultant objects were not joined together when restored to their original sizes. Dilatation followed by erosion could be used to fill in holes that occurred in the digital images of plaques that resulted from tiny opaque particles that were always present in the agar.

Then the program then entered an 'editing cycle' in which it was possible to separate plaques manually by drawing a line between them. Holes in plaques that had not been filled by dilating and eroding could be patched. Most importantly, it was possible to reject objects that were not normal single plaques. I only divided and accepted touching plaques if overlap was very small, *i.e.* a minimal percentage of the area of each plaque was lost by overlap. This was of course subjective but I tried to be consistent. Ideally one might decide to reject all touching plaques but in practice this might mean having to measure extra fields, considerably increasing the time taken, and the use of the apparatus was charged at £40 per day. Occasionally there would

be a plaque very much larger than any other plaque on the plate due a revertant 'phage that had presumably deleted all or part of its palindrome, and this was rejected. Any irregular plaques that looked as though they might represent two close plaques that had merged whilst growing were also rejected, as were bubbles and plaques that partially went off the screen area. On every plate there were plaques that were smaller than the rest, presumed to be due to late adsorption of 'phage onto bacterial surfaces. I decided at the start not to reject these because there was a continuum of plaque size and the problem arose of where to draw the line. This was particularly important because the average size of plaques of different 'phage strains differed considerably and rejecting plaques below a certain size would have a different effect on the overall result for a strain that produced tiny plaques than for a strain that produced large plaques.

This project was about secondary structure in trinucleotide repeat DNA and I could not go into all the details of the population biology of bacteriophage growing on bacterial lawns but it was noticeable that tiny plaques did not seem to be distributed entirely randomly. They appeared to be very close to full-sized plaques more often than expected. In particular, many full-sized plaques had little ones joined to them. The question then arose of whether the little plaques had been independent or whether the large plaque had just grown in a series of bulges as diffusing 'phage met little growing bacterial colonies before the lawn reached confluence. If the little plaque was joined to the large one by a narrow neck, I separated it and accepted both plaques. If the plaques could not be separated because of substantial overlap, the appearance was of a plaque with a bud on it. So many plaques had buds that one could not reject them all. I decided to accept a plaque if it had one small bud - which would make very little difference to the total area - but to reject plaques with two or more buds. When one was happy with all the editing, the field was accepted and the computer estimated the areas of all the individual accepted plaques. One could then move the plate and choose a new field.

Despite titring the 'phage suspensions before plating, some plates would have more or less plaques than planned and in particular might have less plaques than the plate or plates used when choosing the field size. When measuring three fields per plate this sometimes resulted in less than 60 plaques being measured per plate, but I soon started measuring extra fields when required. Measuring multiple fields on a plate, especially if plaques were quite sparse, made it necessary to mark areas already measured to prevent plaques from being measured twice. It was possible to fix an image on the screen and so one could remove the plate from the stage, hold it up to light, identify the region that corresponded to the screen and draw round it on the bottom of the plate with a pen. This, however, was no use if the illumination was such that the background was so dark that one could not see the marks on the screen when the plate was returned to the stage. It was therefore necessary to have a background pale enough that marks could be seen even though this made decision about the position of the edges of plaques a little more difficult.

Initially I chose for measurement three areas on each plate that had relatively high plaque densities so as to measure as many plaques as possible in three fields but I later realized that most plates had slightly higher plaque density and coinciding slightly higher average plaque size on one side of the plate than the other, probably due to a slight gradient in thickness of the top agar across the plate. I then adopted the practice of placing fields symmetrically, *e.g.* one on either side and one in the middle, on a diameter as close as possible to the one that divided the smallest sparsest plaques from the largest most densely-placed plaques, so that fields would not only have similar and average-sized plaques but also similar numbers.

Finally, despite using equal volumes of the same plating-cells on all plates, some lawns were more or less translucent than others. This meant that it was necessary to alter the threshold 'grey-level' for pixel appearance for some plates so that the digital images still corresponded to the edges of the plaques in the same way as on other plates rather than being too large or too small. The above may convey the impression that plaque measurement was so imprecise and subjective that no

meaningful results could be obtained. In fact (a) the plaques produced by different 'phage strains often differed by several times in area and observer error was very small compared with this, and (b) it was not possible to guess the median size in square millimetres of plaques in a field let alone of all the plaques on all the fields of all the plates of a particular 'phage strain so there was no danger of observer bias. I was as consistent and careful as possible to reduce measurement error to a minimum and was able to obtain reproducible results.

Though the PSQ agar was always made up in the same way, and the rest of the methods were also the same, the size of all plaques on all plates could be noticeably different between assays, perhaps due to differences in the rate of drying of the plates at different atmospheric temperatures and humidities, so a reference 'phage, DRL176 (see Materials, Chapter 2) was used so that results from different assays could be compared.

Processing the PSQ data

When each edited field was accepted, the plaque area results were printed out and were stored in a file in Quantimet. All data went into the same file until a command was given to start a new one. It was important to remember to do this after every few plates because the files were of limited size and when full the data was lost and no warning was given. The files could subsequently be copied onto a floppy disc and so transferred to another computer for analysis. Lost data had to be typed in from the hard copy.

The data was analysed using the Microsoft Excel 'package'. The exact number of plaques measured on any plate was determined by the number of acceptable plaques there happened to be in the fields viewed (up to a total of about 120 plaques) and no attempt was made, either at the stage of measurement or at the stage of analysis, to limit the number counted so as to have data on the same number of plaques on every plate as it was felt that to exclude some plaques on the grounds of being surplus could introduce bias. At first I used the same two display and

measurement methods as my predecessors, Angus Davison and Thorsten Allers. The plaques of any one 'phage strain vary in size. To compare the ranges of plaque area of different strains graphically, plotting numbers or proportions of the plaques that have a particular area, a much clearer picture is given by plotting cumulative frequency curves than overlapping bell curves or overlapping histograms. Examples of cumulative frequency curves may be seen in Chapter 4. Secondly, the median plaque area was used in numerical comparisons rather than the mean because the median is less affected by outliers, and the median used was that of all plaques of a 'phage strain or isolate pooled from measurements on different plates. The median does not have a standard deviation or a standard error and reliance on the result that the median plaque area of one strain differed significantly from that of another was based upon measuring large numbers of plaques and by seeing how much or little overlap there was in the ranges of area in cumulative frequency plots.

Modifications to the PSQ protocol

After the initial learning process and the making of the measurements reported in Chapter 4 (and some others), the antiquated Quantimet image analyzer was replaced with a more modern system. At the same time, I ran an experiment to test the assumption that results from different assays could be scaled using the results of the reference 'phage strain (DRL176). The result led on to a series of other investigations of the variables in the method and these, along with experience gained so far and further thought, led to a number of modifications to the protocol to improve the accuracy and precision of the results. Some of this will be described, but first, the measurement system.

Image analysis

The new installation was an Optimas system (Optimas UK, West Malling, Kent) using Optimas 5.2 software with a Visionplus AFG image capture board and a

Pulnix TM - 6 monochrome camera run with twin monitors mounted on a Dell XMT 5100 PC clone. The same camera mounting with stage and light box were used as before. The new system was more 'user-friendly' but much more complicated. A macro called "Plaque" was written by John Findlay (Science Faculty Electron Microscope Facility, University of Edinburgh) which set up appropriate editing boxes on screen. (The provision of two screens meant that the editing options could be chosen on one screen and the plaque images displayed on the other which saved having to remove writing from the screen to see the underlying parts of plaques or having the plaque images in a small window.)

The machine was calibrated in the same way as Quantimet. After the field had been chosen and the grey image 'captured' (fixed on the screen independent of subsequent plate movement) the next operation was automatic background smoothing which helped to eliminate image variation due to imperfect illumination. Then digital image was switched on and was displayed as a red line round the edge of each plaque, the outside edge of the line being the outer limit of the digital image. The procedure I adopted was to mark my estimate of the edge of a plaque on the screen with a fingernail, then to switch on the digital image and see whether the red line just touched my fingernail. I did this for usually three plaques near the centre of the screen and if not satisfied with the result would then alter the threshold and try again. The machine was good at choosing the right threshold and with most sets of plates the threshold did not have to be altered very often, though I always checked every field. Eroding and dilating, filling in holes, cutting and patching, and rejecting plaques could all be done, as on the other machine though in a slightly different way.

Another feature was that when the plaques of a field had been quantified the measurements were all displayed in a box on the control screen. Each plaque could be identified so that its area in mm^2 could be seen. This still did not introduce bias but a number obviously out of keeping with the rest did occasionally draw attention to some aberrant object on the screen that had previously escaped notice. Once any such correction had been made, the results could be transferred directly into an Excel

worksheet in a window on the control screen, labelled and the worksheet resaved. This saved loss of data when the system crashed, which was one of the problems with Quantimet. However, any data that *was* lost due to crashing, or certain 'bugs' in the system, could not be recovered because there was no hard copy. Plates just had to be measured again.

Standardization of results

In order to verify the assumption that the results from one set of plaque area measurements could be compared to those of another by multiplying by the ratio of the median areas obtained for the reference 'phage (DRL176) in the two assays, it was necessary to show that the median areas of the plaques of two different 'phage strains maintained the same ratio to one another under the range of conditions that might apply during plaque area assays. Since great care was taken in making up the media, measuring exactly the same volume of bottom and top agar onto each plate, volume of plating cells *etc.*, it was felt that the main source of variation in plaque size between assays was variation in salt and water concentrations in the agar due to (i) different loss of water during autoclaving and (ii) different rates of drying of the plates over the three days in the plate-pouring room under different atmospheric conditions. It was therefore decided to pour a large batch of PSQ plates and to take a few on each of a series of days and pour on them the same two 'phage strains each time (on N2364 cells with PSQ top agar as for any plaque area assay) and observe the ratio of the plaque areas as the plates became steadily drier and so the plaques smaller. By happy accident another source of variation in plaque size was also introduced which sometimes led to plaques unexpectedly increasing in size.

Plating was carried out on days 2, 3, 4, 5, 6, 8 and 13 after pouring of the bottom agar. The strains chosen were DRL176 and DRL224 and four plates of each were poured on each day except for day 3 when five of each and of a third strain were poured, for a purpose described in Chapter 7. The results are shown in Table 3.1 and

Day	2		3		4		5
λDRL Strain No.	224	176	224	176	224	176	224
Mean	1·8484	1·6258	1·5439	0·5167	1·7700	0·7265	1·1284
Standard Error	0·1132	0·0656	0·0234	0·0083	0·0287	0·0110	0·0234
Median	1·1571	1·8031	1·5993	0·5076	1·8232	0·7375	1·1845
Mode	4·1361	0·3057	1·7013	0·6043	1·5992	0·6507	0·9021
Std. Deviation	1·6827	0·9661	0·4840	0·2116	0·4854	0·2421	0·4387
Minimum	0·0239	0·0919	0·1213	0·0934	0·2319	0·1301	0·0779
Maximum	5·7161	3·5216	2·9878	1·2183	2·7898	1·8979	1·9083
Count	221	217	428	656	286	484	352
95% Confidence lt.	±0·2218	±0·1285	±0·0458	±0·0162	±0·0563	±0·0216	±0·0458
Ratio of medians	0·6417		3·1506		2·4723		

Day	5	6		8		13	
λDRL Strain No.	176	224	176	224	176	224	176
Mean	0·4638	1·0351	0·3850	1·8233	0·7549	0·5499	0·2151
Standard Error	0·0062	0·0208	0·0084	0·0328	0·0090	0·00995	0·0044
Median	0·4672	1·0505	0·3703	1·9379	0·7597	0·5604	0·2148
Mode	0·4049	0·9688	0·2897	2·0322	0·6535	0·5686	0·1829
Std. Deviation	0·1459	0·3348	0·1546	0·5208	0·1600	0·1591	0·0737
Minimum	0·0288	0·0697	0·0034	0·0638	0·2134	0·0396	0·0369
Maximum	0·9691	2·1351	0·9329	2·7024	1·2235	0·9479	0·4897
Count	549	258	335	252	314	256	277
95% Confidence lt.	±0·0122	±0·0408	±0·0166	±0·0643	±0·0177	±0·0195	±0·0087
Ratio of medians	2·5354	2·8370		2·5510		2·6070	

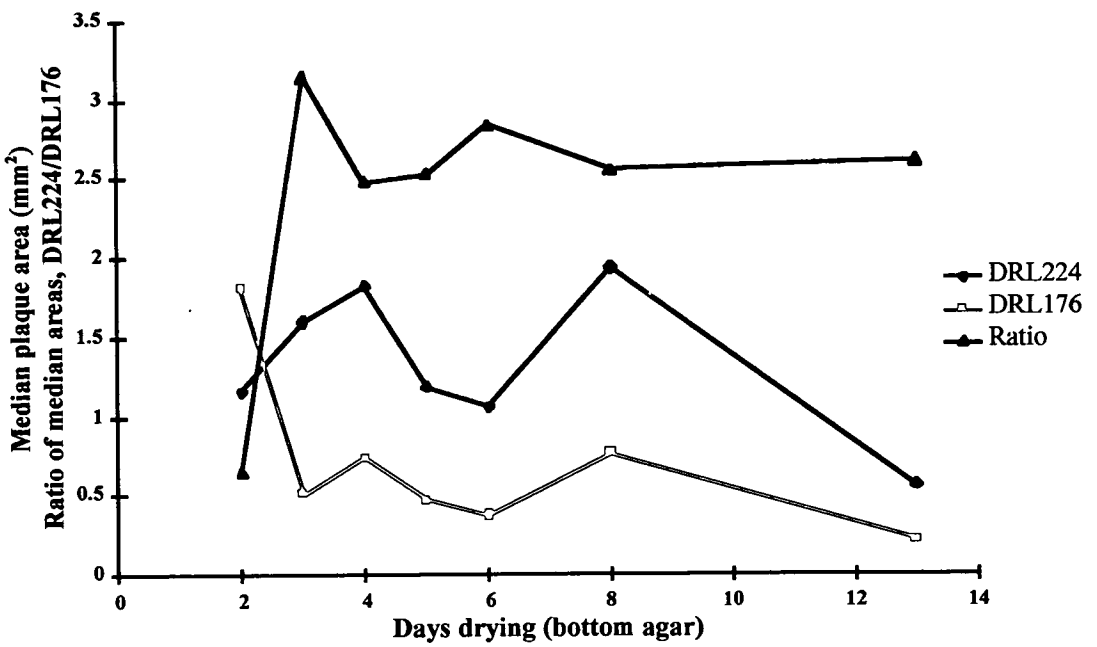


Table 3.1 and Figure 3.1 Plaque area measurements in mm^2 of two bacteriophage strains to determine whether the ratio of median areas remains constant (see text).

Figure 3.1. DRL176 and DRL224 were chosen because they were known to have quite different plaque sizes, DRL224 much larger than DRL176. The results on Day 2 were not expected to give a reliable result because, as was stated earlier, plaques on plates left to dry for less than three days tend to be very large and variable. The appearance on the Day 2 plates was that the plaques of DRL224 were larger than those of DRL176, and the mean and mode of plaque area were larger, but the median was smaller, giving a ratio of < 1.

It was expected that plaque area of both 'phage would decrease exponentially with time, and it was hoped that from Day 3 onwards the ratio would be constant. From Day 3 the plaques of the two strains always varied in the same direction as one-another and from Day 4 onwards the ratio varied little, but on Day 3 it was rather higher and on this basis it was decided that in future it might be better to leave the plates drying for 4 days before use for PSQ. It was notable that the ratio was almost the same on Day 13 as on Day 8 though the plaques were smaller by a factor of almost 3. It was therefore decided that the use of a reference 'phage to scale results from one plaque assay for comparison with those of another was justified, though realizing that this would never be perfect but would probably increase in accuracy the more plates were measured.

Cell growth

The main surprise from the above results was that on two occasions the plaque size had gone up rather than down. The answers lay in the age and condition of the plating cells and the length of incubation of the plates. Normally, for a plaque assay the N2364 plating cells were made up on the day of use from a fresh overnight culture and kept in the refrigerator for a few hours before use, and the plates were incubated overnight for at least 16 hours. This particular assay, however, had been carried out over period of time while much other work was being done and old cells were often used.

Bacteriophage plaque growth on a plate is limited by the 'phage finding itself entirely surrounded by cells that have become contiguous and stopped replicating. Subsequent investigations showed that normally plaque growth has halted before 12 hours of incubation but when the titre of viable plating-cells is lower than normal, either because the overnight culture was very old or because the plating cell culture is old, the viable cells on the plate will take longer to become contiguous and the plaques will be larger if incubated long enough. It was also found that plaques would be larger if the plating cells, made on the day, had been kept at 4°C for a few hours rather than being used immediately, while still warm. On Day 3 a new overnight culture was used to make a plating cell culture on the day. The plating cells were made up in the evening and used soon afterwards. On Day 4 the same plating cell culture was used, having been kept at 4°C for about 23 hours, and the plaques were larger despite the plates having dried for an extra day. The time of incubation of the plates was about 17½ hr each time. The effect was probably mainly due to the fact that the cells had been kept cold rather than that they were older because the same cells were used again on Day 5 yet the plaques were about 35% smaller than on Day 4, no doubt due mainly to further drying of the plates. (The incubation time was 16 hr 30 min.)

The plating cell culture used on Day 8 had been made 3 days before with an overnight culture that had been taken out of incubation two days before that (and kept at 4°C), and on Day 9 I forgot to take the (Day 8) plates out of the incubator and they remained at 37°C for 47 hr 49 min. On Day 13 the very same plating-cell culture, now 5 days older, was used, but the plates were only incubated for 16 hr 0 min. The plaques on such dry plates were tiny, but probably would have been even smaller had the plating cells been new.

From these observations, two changes to the protocol were decided. First, it was decided that in future not only should I be scrupulous about making up fresh plating cells for plaque size quantification assays, as indeed I always had been before this, but that the cells should be made up just before use and, if required to be kept

waiting at all, then this should be on the bench at room temperature rather than in the refrigerator. Secondly, till this time, plates had been taken out of the incubator and taken straight to the image analyzer for measurement. Quantimet generated a lot of heat and was in a small room with no windows so the temperature was often high. Probably the plaques had finished growing because fresh plating-cells were used, but if they had not, they might go on growing in the image analysis room so that plaques measured at the end of the day might be larger than ones measured at the beginning. The plates were then kept at 4°C overnight and measurement of remaining plates was continued next day. It was therefore decided that in future (a) overnight incubation would always be exactly the same period, 16 hr, and (b) plates would then be put into the 4°C room and measurement not started until the next day.

Bottom agar volume

Another variable investigated was the volume of bottom agar. 44 ml was about the maximum that could be used because there had to be room for the top agar and to tip the plate to spread the latter without it overflowing. It was decided to measure the plaque areas with one strain (DRL176) on five plates each of 36, 38, 40, 42 and 44 ml bottom agar (with the usual 2·5 ml top agar on all). In order to arrange that, as usual, sets of plates being compared all had members from different parts of the stack, plates were labelled before pouring with the volume they were to hold. Then the first plate was poured with 36 ml, the next with 38 ml, then 40, 42, 44, 36, 38 ml *etc.* up the stack. Enough plates had to be poured so that if any plate grew a contaminant during the four days drying, it could be replaced with another plate of the same volume. As it happened, none was infected and the 25 plates above the bottom two were used. The results are shown in Figure 3.2, overleaf.

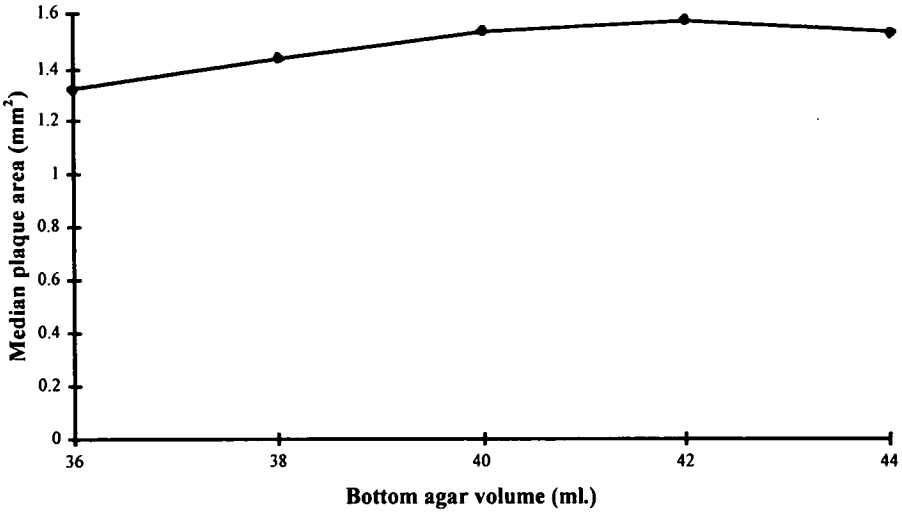


Figure 3.2 Effect of volume of bottom agar on plaque size.

From these it was decided that in future 42 ml might be a better volume to use because any slight errors that might be made in pouring would make less difference to plaque size than at 40 ml. There is however another reason for relating this little experiment and showing this rather boring graph.

The plate position effect

No series of plates was poured specifically to examine the effect on plaque size of differential drying of plates in different parts of the stack but information about this could be gained from examination of the results obtained from plates poured for other purposes. Until the time of these investigations, plates had not been numbered before use with their position in the stack but it was possible to reconstruct this information because meticulous notes were made of what was done and because plates were all labelled not only with the working number of the 'phage construct and the isolate number but also with the plate number for that isolate. Thus if, as was usually the case, plates were dealt from the top of the stack into N piles, one for each isolate, and 5 plates in each pile, then the last plate put down on the top of each pile would be numbered '5' and the first, at the bottom, '1'. This

number was always entered in the labelling of the data set for every field of every plate so that if there was ever any doubt about any result the exact plate could be identified and re-examined. The stack of plates for the investigation of the effect of bottom agar volume had 37 plates and the plates used were numbers 11 - 35. The median plaque areas for the individual plates are plotted in Figure 3.3a with lines connecting plates of the same volume. Figure 3.3b shows the results transformed by multiplying the median of each non-42-ml plate by the ratio of the median of all plaques on 42 ml plates to the median of all plaques on plates of the volume concerned, thus 'normalizing' all results to 42 ml. Then the lines through the points have been smoothed. There is obviously some residual variation not due to position of plates in the stack but it is quite clear that plaques from plates near the centre are larger than those from plates near the bottom. There is a suggestion that the size might go up again a little right at the bottom but it seemed that this would probably not be great enough to warrant the usual practice of discarding last two plates (as had been done in this series).

Another series of plates was examined to look at the position effect at the top end of stacks. It was one of several series of plates poured to examine the ratios between the areas of the plaques of different 'phage strains at different incubation times (data not shown). The plaques had stopped growing before the shortest incubation time used (20 hr), *i.e.* the plaques of any one strain were the same size at all incubation times used, making the data quite suitable for plate position effect examination. 15 plates were poured of each of three strains and the usual practice had been followed of using all plates in the first stack bar the first two and the last two and then going on to the second stack. Figure 3.4 shows the plaque areas, in this case with each line connecting all the plates of one stack. This time there does not appear to have been much difference between plates in the middle and at the bottom of the stacks but, at least in stack 2, plaque size does seem to be smaller in plates near the top, where one normally finds the most drying to occur. On the basis of the results in Figures 3.3 and 3.4 it was decided that if four plates of a stack were to be

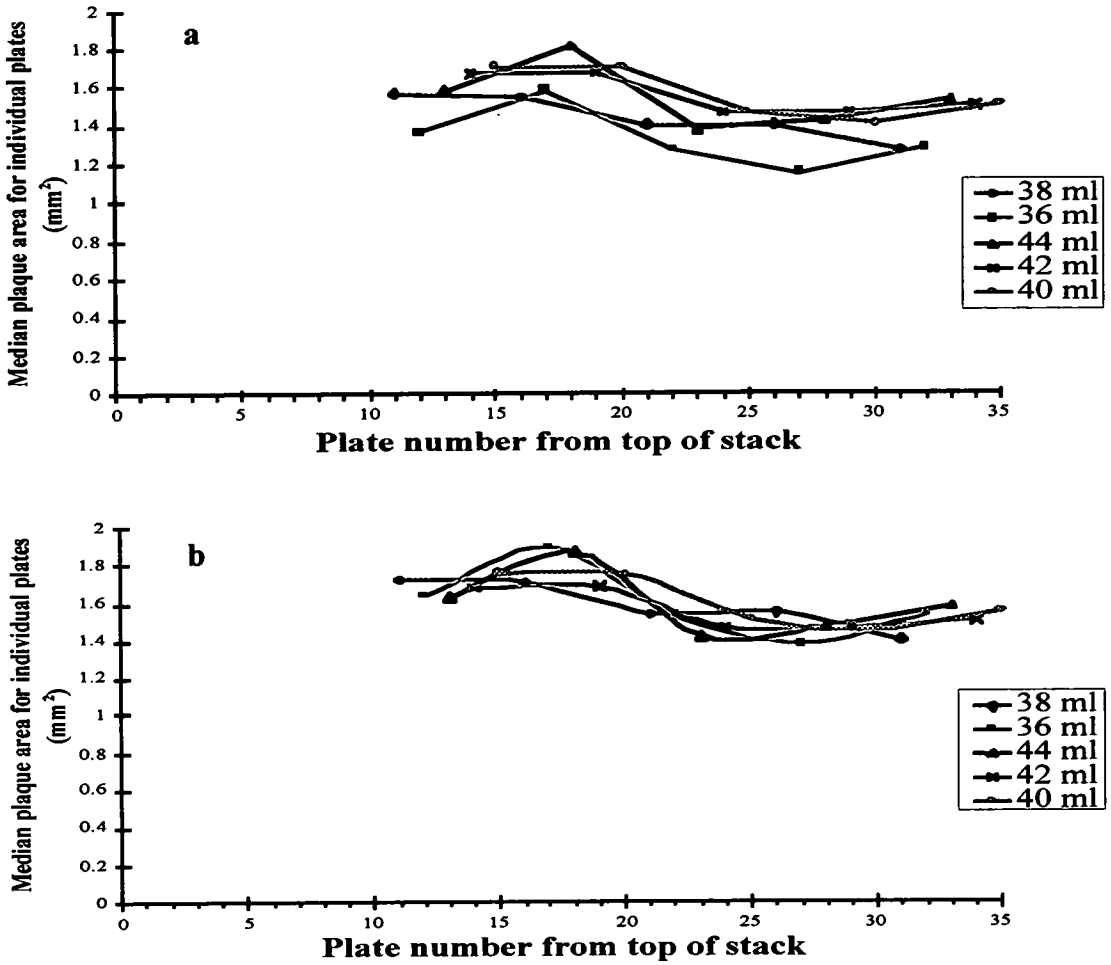


Figure 3.3 Median plaque areas of individual plates of the assay shown in Figure 2.2 plotted against the positions in the stack in which the plates had dried. (a) Raw data (b) Data normalized to a volume of 42 ml and curves smoothed.

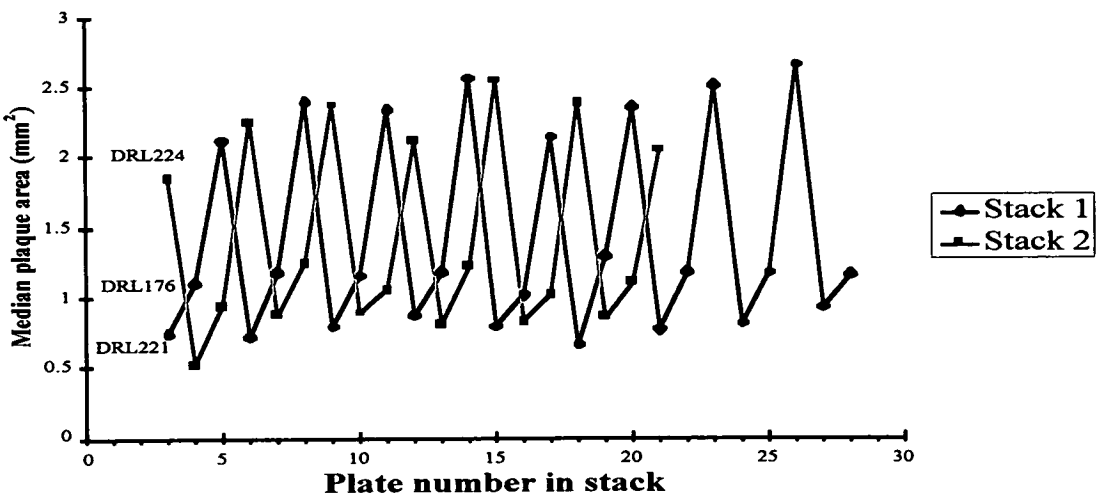


Figure 3.4 Median plaque areas of 3 'phage strains on plates from two stacks.

excluded from use for plaque size quantification it would be better to exclude the top four rather than the top two and the bottom two. The results in Figure 3.4 also suggested another change to protocol. The Stacks each contained 30 plates. Removing the two at each end left 26. This meant that dealing plates in rotation to three piles gave eight complete rounds. On the ninth round, two 'phage strains received a plate from the bottom of the first stack and the other had one from the top of the next stack. Since plaques tend to be of different sizes on plates from opposite ends of a stack and sometimes, as will be seen from data to be presented later, the plaques from all plates in one stack are smaller than from plates in the same positions in another stack, it was decided that in future only a whole number of rounds should be dealt from any one stack and each strain should have an equal number of plates from each of the stacks.

Usually six strains were to be compared in any one assay so it was decided to use eight plates for each strain, making a total of 48 plates used (which is about as many as could practically be handled at once) and that these should be distributed from two stacks, four plates to go to each strain from each stack. Since 12 plates could be poured from one bottle of bottom agar, five bottles would need to be poured.

The revised protocol

Five bottles of PSQ agar were made up with the increased quantities (see Materials, Chapter 2) to give volumes of roughly 525 ml to allow the pouring of twelve 42-ml plates from each bottle with some spare volume to provide for evaporative loss in autoclaving. (There is enough room to fit 525 ml in a 500 ml bottle leaving some air-space.) Two 100 ml bottles of PSQ bottom agar were made up. All were autoclaved together. Then the top agar was stored at 60°C and the bottom agar was cooled to 46°C and two stacks each of 30 × 42 ml plates were poured on a levelled table and left to set overnight. Next day, without disturbing the

positions of the plates relative to one-another, every plate was labelled on its side-wall with its stack number and its number in the stack, from the top. The plates were then left to complete four days of drying at room temperature.

In the meantime plate lysates of the 'phage strains to be compared were diluted and plated as before. Two isolates of each strain were plated if two that showed evidence of having an insert were available. Then *one* plaque was picked from each plate and plaque suspensions made and diluted as before. Then *50* µl of each dilute plaque suspension was plated for titring. Previously five plaques had been picked after the initial plating and five separate suspensions made, one for each PSQ plate that was to be poured, and 10 µl of each was plated for titring. These titres proved to be not very reliable for estimating the volumes required to plate 350 p.f.u. on each PSQ plate, especially for older lysates, and it was felt better to plate larger volumes. 50 µl gave about 250 - 1050 plaques per plate from fresh lysates. Counting these numbers of plaques on 48 titre plates would take a long time but since 'phage had been plaque-purified before the lysates were made there was no necessity to make suspensions from separate plaques for each PSQ plate hence making just one suspension for each lysate. Then these large numbers of plaques would have to be counted on usually at most only 11 plates (as a single suspension of the reference 'phage, DRL176, was used for pouring the eight reference plates, four in each stack).

Under the old protocol, plates were usually poured on a Monday and were then three days old on the Thursday when the 'phage were plated, and plaque measurement began the next morning. With the new protocol of four-day drying, plates were usually poured on a Thursday so that they were four days old on a Monday. Therefore an overnight culture of N2364 was set up on the same Thursday that the plates were poured, taken out of the incubator on the Friday morning and put into the refrigerator over the weekend to use for making plating cells on the Monday. So, if plates were poured on different days from these, the overnight culture was still set up on the same day as the plates were poured so that it was the same standard age before use. Making and titring of the 'phage suspensions was

usually done during the week that the bottom agar was poured. The titres of 'phage suspensions were found to drift down if exposed to the constant artificial light in the cold-room but would keep for long periods at 4°C with very little change in titre if covered. (A hood of aluminium foil over the tube-rack was used.)

On the day of plating the 'phage, three plating-cell cultures (two for use and one as a spare) were put into incubation to be ready at about the time they would be required and if there was any delay they were then kept at room temperature until used. The stacks of plates were examined to see whether any plates were infected. If not, the top four plates of the first stack were set aside and then the next 24 (plates 5 - 28 of the stack) were dealt in rotation into six piles of four plates each and each plate labelled on its edge with its strain and isolate number and the number of the plate in that pile (4, 3, 2, 1 from the last plate put down). The same was done with the second stack, labelling the plates for the second isolates of the same 'phage in the same order. (Plates for DRL176 and any 'phage for which there was only one positive isolate, would be labelled 5 - 8, *i.e.* 8, 7, 6, 5).

If any plate was found to be contaminated, the same plate was removed from the other stack and dealing continued down to plate 29 of each stack. If any other plate or plates showed growth, the bottom plates were used before resorting to the use of plates 4 or 3 of the stacks. Occasionally several plates were randomly contaminated in one stack and only one or none in the other and then a scheme of giving plates to the shorter stack from the longer one would be worked out to give as close as possible equivalence between the stacks. Table 3.2 shows an example. In this case, the contaminated plates were numbers 14, 18, 22, 27 and 30 in Stack 1 and number 27 in Stack 2. By giving two plates from Stack 2 to the remaining set of Stack 1, matching 'phage isolates were able to have roughly equivalent plates. With more thought it might have been possible to devise a slightly better scheme but time also comes into the equation when there is much to be done. As in Tables 2.1 and 2.2, the figure in outline is the number of the 'phage construct and that in normal text the isolate number. In this particular assay six new constructs were to be compared,

	Isolate plate	37,7	40,61	41,29	44,1	45,2	46,4	Cell bottle	Top agar bottle
Isolate	1	4_1	5_1	6_1	7_1	8_1	9_1	1	1
1	2	10_1	11_1	12_1	13_1	15_1	16_1	2	1
Mainly	3	17_1	17_2	19_1	20_1	21_1	23_1	1	2
Stack 1	4	24_1	25_1	26_1	28_1	29_1	29_2	2	2
Isolate	1	4_2	5_2	6_2	7_2	8_2	9_2	1	1
2	2	10_2	11_2	12_2	13_2	14_2	15_2	2	1
	3	16_2	18_2	19_2	20_2	21_2	22_2	1	2
Stack 2	4	23_2	24_2	25_2	26_2	28_2	30_2	2	2
		39,9	40,73	41,34	44,7	45,3	46,19		

Table 3.2 Illustration of a scheme used to compensate for the loss of contaminated plates in one stack to be used for a plaque assay. Numbers in the body of the table are plate numbers counting from the top of the stack and subscripts are stack numbers.

---

leaving no room for the usual reference 'phage, DRL176. However 37,7 and 39,9, which were thought at the time to be the same, had been measured in a previous assay so they were used as the plaque area reference in this case.

When the plates were all dealt and labelled they were rearranged in the order that they were to be plated. Plates 1 and 2 of the first isolate of the first test strain were placed on top of plates 1 and 2 of the second isolate of the same strain and they on top of plates 1 and 2 of the first and second isolates of the second and third strains in succession. Plates 3 and 4 of the first isolate of the first test strain were placed on top of plates 3 and 4 of the second isolate of the same strain and they on top of plates 3 and 4 of the first and second isolates of the second strain and so on. Then 6 rows of 4 tubes were made in each of two racks. In the first row of the first

rack they were labelled 1 - 4 for the first isolate of the first 'phage strain, followed by 1 - 4 for the second isolate of the same strain, then 1 - 4 of the first isolate of the second strain and so on with usually the last two rows of the second rack labelled 1 - 4 and 5 - 8 for the reference strain DRL176. The volume of one plating cell culture was 10 ml from which 38 or 39 volumes of 250 µl could be dispensed, *i.e.* not as many as 48. Exactly the same was true for portioning volumes of 2·5 ml of top agar out of 100 ml bottles. So, cells, from one bottle were measured into tubes 1 and 3 of every row of both racks and cells from another bottle into tubes 2 and 4 of every row. Then the calculated volumes of the dilute 'phage suspensions to give 350 plaques were added to each of their respectively labelled tubes of cells and the tubes incubated at 37°C for 15 min.

In the meantime a specially-made large pouring table was levelled on the bench and the bottles of top agar were put into a water-bath at 46 - 50°C. Bottom agar had always been poured on a plate-pouring table but previously top agar had been poured on plates on the laboratory bench. The table was introduced to try to minimize differences in plaque size and density across plates due to difference in top agar thickness. Then one of the bottles of top agar was opened and the 24 plates labelled 1 and 2 were poured from the tubes labelled 1 and 2, all the plates having been arranged in the same order as the tubes. Then the other bottle was opened and all the plates labelled 3 and 4 were poured from their respective tubes. The use of a little round spirit level showed that, as Physics dictates, no table is level all over; it always sags in the middle. The table I used measures 120 cm × 2 ft and has six adjustable legs. It was adjusted so that there was a level gutter along its long axis from which it sloped up slightly to all sides and corners. The plates were poured at the right-hand end of the level area and moved gently as far left as possible within it immediately and left to set. The level area can accommodate eight plates with a little room to spare (and one has to have the Bunsen burner nearby) and by the time it was full, plates at the left end were set and could be moved into be lines in other areas.

When all poured and set, the plates were put into the 37°C incubator. One whole shelf of the incubator was occupied by the set of plates. It could take three rows of four piles of four plates. The incubator was used by many people and often opened so was no doubt often cooler at the front than at the back. To compensate for any difference that might be caused by this, two plates of each isolate of each of the first four test 'phage were put in the back row. All the plates of the remaining two strains, one of which was the reference 'phage, were put in the central row and the other two plates of each isolate of the first four 'phage were put in the front row. In addition, the plates were usually poured in the evening rather than the afternoon. This meant that the plates were disturbed very little in the initial hours of growth but mainly in the last few of the exactly 16 hours incubation when there would be little remaining growth. After removal from the incubator the plates were usually inspected briefly and then were put into the cold-room at 4°C until the next day.

Modification of the analysis

The above modifications to the method undoubtedly improved consistency, but there was still between-plate variation in plaque area for any strain. There was less variation in plaque numbers per plate but there was still some. The median of all plaques measured of one strain could therefore be skewed by a plate simply because it happened to have more plaques. To avoid this, the median plaque size was determined for each plate and, taking this as the best estimate for the plate, these medians were treated as single data points and the mean was calculated of all eight results (having first established that both isolates of a phage behaved the same way). This meant that it was possible also to give a 95% confidence range for the mean of plate medians. The calculated range would be a little less than the real range since the medians would not be perfect measures of the central plaque area of each plate but with so many plaques measured they would be quite accurate and previously no confidence limits had been expressed at all. After calculating the plate medians and

before calculating the means, six histograms were displayed showing the eight plate medians of each of the six strains compared. This not only immediately showed whether one isolate was producing different plaques from its fellow but gave six profiles of plaque variation through each of the two stacks. The overall impression was that all phage were affected in the same way by the vagaries of the rates of drying in different parts of the stacks.

Chapter 4

d(CAG)·d(CTG) and d(GAC)·d(GTC) repeats

Introduction

The work described in this chapter investigated the potential of d(CAG)·d(CTG) and d(GAC)·d(GTC) repeats to form hairpins *in vivo*. During the year in which it was published (1995) several reports appeared of *in vitro* investigations of the same.

Mitas *et al.* (1995a) divided trinucleotide repeats into four classes: Class I which are (G + C)-rich and contain GpC or CpG dinucleotide palindromes, *i.e.* $d[(CGG)\cdot(CCG)]_n$, $d[(CAG)\cdot(CTG)]_n$ and $d[(GAC)\cdot(GTC)]_n$; Class II, which are (G + C)-rich but do not contain GpC or CpG dinucleotide palindromes, *i.e.* $d[(CAC)\cdot(GTG)]_n$ and $d[(CTC)\cdot(GAG)]_n$; Class III, which are (A + T)-rich and contain ApT or TpA dinucleotide palindromes, *i.e.* $d[(ATC)\cdot(GAT)]_n$, $d[(ACT)\cdot(AGT)]_n$ and $d[(TAT)\cdot(ATA)]_n$ and Class VI which are (A + T)-rich but do not contain ApT or TpA dinucleotide palindromes, *i.e.* $d[(TGT)\cdot(ACA)]_n$ and $d[(TCT)\cdot(AGA)]_n$. (They also named a Class V of trinucleotide repeats but these were actually mononucleotide repeats.) They pointed out that all the disorders then known to be due to trinucleotide repeat expansion were caused by repeats in Class I and observed that these could potentially form hairpins, though, in this paper, they only considered the possibility of hairpins containing an odd number of trinucleotides (see p. 71). Class III repeats might also form hairpins but these would be expected to be less stable since A·T base pairs are less stable than G·C pairs.

Mitas *et al.* (1995a) computed theoretical energy minimizations of hairpins formed by different odd numbers of repeat units for each of the six single strands of their Class I repeats (These did not include the possibility of T·T or G·G bonds *etc.*.) They concluded that the order of stability was CTG>CCG>GTC>

CGG>GAC>CAG and that the difference in stabilities of $d(CTG)_n$ and $d(GTC)_n$ hairpins arose from stacking energies. They showed that oligonucleotides containing $d(CTG)_{15}$ or $d(CAG)_{15}$ had electrophoretic migration rates the same as palindromic sequences of the same length, migrating more rapidly than double-stranded DNA, whereas oligonucleotides containing repeats not expected to form hairpins [$d(ATC)_{15}$ and $d(GAT)_{15}$], and random sequences, had migration rates less than that of double-stranded DNA. They also found that all the thymines in the sequence $d(CTG)_{15}$ were protected from oxidation by $KMnO_4$ except the central one of the repeat sequence, and that P1 nuclease also cleaved the oligonucleotide in the predicted loop region and not elsewhere and so concluded that a hairpin was formed containing all the repeat units and that the thymines within the stem must be involved in base-pairs. They then computed an energy minimization for $d(CTG)_{15}$ including T·T bonds and found that these bonds hardly added to the stability at all and so concluded that they must have other deleterious effects.

Subsequently the same workers (Yu *et al.*, 1995a) carried out similar investigations on an oligonucleotide containing $d(GTC)_{15}$ and compared it with the one containing $d(CTG)_{15}$ with the same flanking sequence. The one containing $d(GTC)_{15}$ also formed a hairpin involving all fifteen trinucleotides of the repeat sequence. Cleavage with different concentrations of $KMnO_4$ at different temperatures in 50 mM NaCl with or without 150 mM KCl showed that both were more stable in the higher salt concentration but that the $d(GTC)_{15}$ hairpin became susceptible to modification at a lower temperature than did the $d(CTG)_{15}$ one. Electrophoretic mobility melting profiles showed that the T_m of the $d(GTC)_{15}$ hairpin was only about 38°C as against about 48°C for the $d(CTG)_{15}$ hairpin. Cleavage of the $d(CTG)_{15}$ hairpin by P1 nuclease was shown to be at the GpC phosphodiester bond on the 5' side the central (*i.e.* eighth) CTG trinucleotide, not at the CpT and TpG of this central trinucleotide as expected. Cleavage of the $d(GTC)_{15}$ hairpin was detected at four positions in sharply descending frequency in a 5'→3' direction, the most frequently cleaved position being the TpC of the seventh GTC trinucleotide.

The possibility that the oligonucleotides formed hairpins with an even number of trinucleotides (*i.e.* 14) was considered but the cleavage position in the flanking sequence refuted this.

The theoretical structures of these hairpins were then studied by computer modelling. This suggested that the unpaired loops, especially the GTC one, bend over in such a way that the bonds cleaved by P1 are more exposed. Yu *et al.* (1995a) concluded that this was due to the nature of guanine-guanine stacking. They went on to mention that a $d(CGG)_n$ single strand has the possibility of forming a hairpin with d(CGG)·d(CGG) pairing in the stem and an odd number of bases in the loop or of d(GGC)·d(GGC) pairing in the stem (or, as they put it, CGG pairing with GCG) and an even number of bases in the loop. (Actually there are four more possibilities as will be discussed in the next chapter.) They then suggested that if the most stable loop structure of the $d(CGG)_n$ hairpin is not compatible with the most stable stem structure a flexible hairpin might result. Strangely, however, they did not extend this to wonder whether their results with their $d(CTG)_{15}$ and $d(GTC)_{15}$ oligonucleotides might not have been caused by a conflict between an even-membered loop possibly being more stable but the DNA being constrained to pair with an odd-membered loop because of the odd number of repeat units and the flanking bases.

Investigation of oligonucleotides containing $d(CAG)_{15}$ and $d(GAC)_{15}$ (Yu *et al.*, 1995b) indicated that they too both formed hairpins with odd-membered loops. Each cleaved with P1 nuclease at three points within the predicted loop. The T_m of the $d(CAG)_{15}$ hairpin was 38°C while that of the $d(GAC)_{15}$ was 49°C and the authors hypothesized that there were A·A bonds in both but that these were stronger in the $d(GAC)_{15}$ hairpin. It might be noted that their T_m results gave an order of stability GAC≈CTG>CAG=GTC which was rather different from the one they had reached theoretically Mitas *et al.* (1995a), CTG>GTC>GAC>CAG.

Gacy *et al.* (1995) studied $d(CAG)_{25}$ and $d(CTG)_{25}$ oligonucleotides. They investigated their absorbances in solution at 260 nm and found marked increases with temperature suggesting melting of hydrogen-bonded structures. The T_m was found to

be 52°C for $d(CTG)_{25}$ and 50°C for $d(CAG)_{25}$. This difference was much less than the 10°C difference determined by Yu *et al.* (1995a,b) for $d(CTG)_{15}$ and $d(CAG)_{15}$ but the respective results for $d(CTG)_{15}$ and $d(CTG)_{25}$ were very similar. Gacy *et al.* (1995) also examined their oligonucleotides in solution by NMR. This too indicated hydrogen bonding and showed features suggestive of a loop structure. Taken together with the absorbance findings these results indicated that the oligonucleotides existed in hydrogen-bonded states consistent with hairpins. To distinguish this possibility from homoduplex formation, the observations were repeated at several DNA concentrations and the results were the same, indicating that the structures were unimolecular. 2D spectroscopy detecting the imino-protons of the thymine residues of the $d(CTG)_{25}$ oligonucleotide determined that these residues were highly stacked within the stem of the hairpin.

In order to determine whether the threshold lengths of trinucleotide repeats associated with expansion could be due to hairpin structures, Gacy *et al.* (1995) edited the sequence files for the d(CAG)·d(CTG)-repeat-containing genes responsible for HD, SCA1 and DM (and the d(CGG)·d(CCG) repeat of *FMR1*) to contain the reported upper limits of the respective normal ranges of repeat lengths and processed them with an RNA folding program which they modified to simulate DNA at 37°C. Examination with a magnifying glass of their grossly overshrunk figure of the structures they obtained shows that it is illegible but it is clear that (a) they only considered structures formed by the coding strands, *i.e.* the strands containing d(CAG) repeats for HD and SCA1, d(CTG) repeats for DM (and d(CGG) repeats for FRAXA), and that (b) they did not cater for the possibility of hydrogen bonding between mispaired nucleotides. The hairpins drawn all included pairing of flanking sequences on either side of the repeat tract and/or pairing of flanking sequence with repeats at the base of the hairpin. The estimated energies were similar, −41 to −53·9 kcal per mole.

Taking this range as a threshold energy necessary for expansion, Gacy *et al.* (1995) used their DNA folding program to investigate the ability of "all 16 classes"

of dinucleotide and trinucleotide repeat to form hairpins to see which could exceed the threshold. Actually the ten different types of double-stranded trinucleotide repeat have twenty different single strands (see pp. 16-17) and the four different dinucleotide repeats have a total of six different single strands (see p. 26). It appears that the authors treated the complementary dinucleotide strands as different but treated the complementary trinucleotide strands as the same and that their DNA folding program was unable to distinguish between the energetic differences of hairpins with for example mismatched adenines and mismatched thymines. However, the authors concluded that only six of their sixteen classes were capable of forming hairpins above threshold energy for expansion - d(CAG)·d(CTG), d(CGG)·d(CCG), d(AT)·d(AT), d(GAC)·d(GTC), d(GC)·d(GC) and d(GT)·d(AC) - and this by their reckoning would seem to be seven classes. The first three were associated with large expansions. The next two, they concluded, were not found to expand because they are not found in long enough stretches to reach the threshold - the longest stretch of d(GAC)·d(GTC) repeats they found in a human genome database was 5 units[3] - and the last, they noted, makes small increases in colon cancer but they estimated that it required 130 repeats to reach the threshold for large expansions and again was not found in sufficient lengths.

In order to confirm that their computer predictions of hairpin stability reflected reality, Gacy *et al.* (1995) melted representative oligonucleotides that they predicted to form hairpins above threshold stability [$d(CGG)_{25}$, $d(CAG)_{25}$, $d(GAC)_5$ and 25 and $d(GT)_{37}$], hairpins of low stability [$d(AAT)_{15}$, $d(ATC)_{14}$ and $d(ACT)_6$] or no secondary structure at all [$d(AAG)_{25}$ and $d(AC)_{37}$] and were only able to obtain T_m results by their spectrophotometric method for three of these, 48°C for $d(CAG)_{25}$ (though stated earlier in the paper as 50°C), 54°C for $d(GAC)_{25}$ and 76°C for $d(CGG)_{25}$. The buffer in which the oligonucleotides were allowed to anneal (at 15°C for 30 min), before measuring absorbance during heating, contained 100 mM NaCl but

[3] Mitas *et al.* (1995a) had found five instances of at least 8 perfect repeats of d(GAC)·d(GTC) but none of these was human.

no K^+. From the diagrams in this paper, it is clear that the authors had noticed the possibility of even- or odd-membered loops to the hairpins but they did not consider whether these might have different stabilities. At the sequence lengths they considered, this may not have been important if all the repeats in a tract formed a single hairpin..

Like Mitas and colleagues (Mitas *et al.*, 1995a; Yu *et al.*, 1995a; Yu *et al.*, 1995b), Mitchell *et al.* (1995) used electrophoretic mobility and chemical modification and cleavage to study d(CXG) repeats but used much shorter oligonucleotides. They made $d(CAG)_5$, $d(CGG)_5$, $d(CCG)_5$, a series of nine trinucleotide repeat oligonucleotides increasing in length by one nucleotide from $d(CTG)_4$ up to $d(CTG)_6CT$, and for controls, $d(GTC)_5$, $d(TCA)_5$ and a 15-mer with a random sequence with the same base composition as $d(CTG)_5$.

All of the d(CXG) oligonucleotides had higher electrophoretic mobilities than the controls, indicating that there was some secondary structure, and all except the controls had one or two faster or slower minor bands in addition to the major band. In the series of d(CTG) oligonucleotides, those of length 12 - 15 nt migrated about 2 nt more rapidly than expected for their length and then there was an abrupt change and those of length 16 -19 nt migrated 3 - 4 nt more rapidly than expected. This step was eliminated by running in a native (polyacrylamide) gel at 60°C but was not quite eliminated by running on a denaturing gel at 20°C though the authors said it was. The figure of the hot native gel shows that at least five of the oligonucleotides still showed minor bands, and some can be seen on the denaturing gel too, but the authors did not discuss the possibility that their oligonucleotide preparations might not be pure.

Melting-point determinations were tried on $d(CTG)_5$ and the random 15-mer but the changes in optical densities were so small that nothing could be judged.

Chemical modification with OsO_4-dipyridine, for T, dimethylsulphate (DMS) for G or hydrazine for C and T residues was carried out on $d(CTG)_5$ and $d(GTC)_5$ at 20°C and at 60°C followed by cleavage with piperidine and electrophoresis. No evidence of secondary structure was seen in $d(GTC)_5$. In $d(CTG)_5$, the only sign of

peripheral bases being less vulnerable to modification than central ones was that the second guanine from the 5' end was not cleaved at 20°C but was at 60°C. The most 5' guanine was apparently not well detected (and the bottom of the gel was cut off the photograph). The authors felt that because the central T was not more cleaved than the others a hairpin was ruled out. However, this was a short oligonucleotide and a hairpin would have had a low melting point; had the authors tried 4°C instead of 20°C they might have had more positive results.

Mitchell *et al.* (1995) then phosphorylated the 5' end of the $d(CTG)_5$ oligonucleotide and tried self-ligation, expecting to see a ladder of bands on electrophoresis, but only obtained one band of reduced mobility. This they imagined was a circle, made possible by juxtaposition of the 5' and 3' ends by the unknown secondary structure.

Smith *et al.* (1995) carried out a very much more detailed NMR investigation of homostrand pairing than did Gacy *et al.* (1995) but with very short oligonucleotides. At 0°C the ^1H imino-proton spectrum of $d(CAG)_2$ indicated random structure but $d(CTG)_2$ showed resonances indicative of antiparallel duplex formation. The ^1H imino-proton spectra of $d(CAG)_3$ and $d(CTG)_3$ were the same as those obtained by Gacy *et al.* (1995) for the 25-repeat strands, indicating that both formed antiparallel duplexes as would be seen in the stem of a hairpin. Smith *et al.* (1995) observed these spectra and that of $d[(CAG)_3{\cdot}(CTG)_3]$ through a range of temperature and noted that though in the heteroduplex the A·T bond signal diminished well before the C·G bond signal ('premelting'), the T·T signal in $d[(CTG)_3]_2$ was as strong as the C·G signal at all temperatures. Melting points were determined both by change in NMR and by change in UV absorbance with temperature. For $d[(CAG)_3]_2$ the estimates were $T_{m.NMR}$ 15°C and $T_{m.UV}$ 23 - 27°C but the latter was thought to be too high because the transition was very broad. The respective results for the other complexes were 27 and 17°C for $d[(CTG)_3]_2$ and 52 and 51°C for $d[(CAG)_3{\cdot}(CTG)_3]$ and it was concluded that the order of stability was

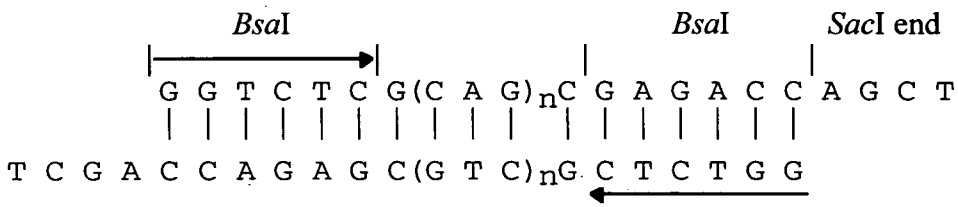
$d[(CAG)_3]_2 < d[(CTG)_3]_2 << d[(CAG)_3 \cdot (CTG)_3]$ and that the presence of A-A mismatches has a more profound effect than that of T-T mismatches.

As with temperature, the T·T signal was maintained through a range of pH, up to 8·5. This could either be due to tight T·T bonds or restricted access of the T imino protons to the solvent water molecules. The authors pointed out that the first possibility was unlikely because the $d(CTG)_3$ duplex was much less stable than the complementary duplex. Resonances due to stacking interactions showed that the thymine residues were stacked into the helix though the pattern was not quite the same as for Watson-Crick base-pairs. Two different T·T pairings can be drawn that both have two hydrogen bonds involving the imino protons of both residues but only one signal was observed. This could be because the two were indistinguishable because of symmetry or because both resonances were degenerate and so overlapped. Other resonances indicated that the arrangement of the sugar-phosphate backbone was not very much disturbed but was compatible with the thymine residues being in fast exchange between the two possible 2-bond pairings. Computer simulation suggested that there were indeed two hydrogen bonds and that the grooves were correspondingly narrowed at the T·T pairs.

Thus all of this work, except for the inconclusive investigations of Mitchell *et al.* (1995), indicated that single strands of d(CAG)·d(CTG) repeats can form imperfect hairpins *in vitro*. As outlined in Chapter 1, the work described here, testing whether this might happen *in vivo*, involved the construction of bacteriophage bearing these repeats in the centre of a long palindrome. The effects of central inserts with different numbers of d(CAG)·d(CTG) repeats are compared with those of d(GAC)·d(GTC) repeats since these have the same bases but have not been found in long arrays, let alone to be unstable. Comparison is also made with sequences of known *in vitro* loop structure and the effect of immediate flanking sequence is checked.

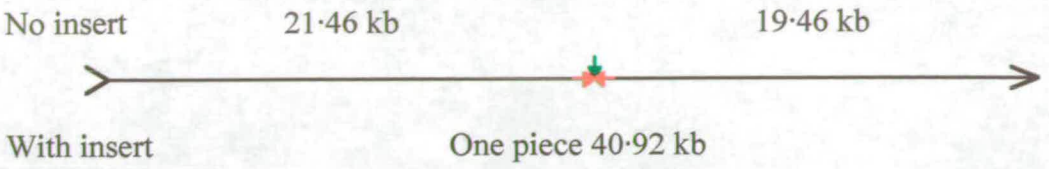
Bacteriophage design and testing

The parent bacteriophage used, λDRL167, has a 462 bp perfect palindrome with a unique *Sac*I site at the centre. The sequence of the palindrome is given in Appendix 1. The inserts containing the test sequences were made by annealing complementary oligonucleotides. The rest of the insert was designed to destroy the *Sac*I site and provide a new restriction site for identification of successful ligation products. It was decided that the new restriction site should be non-palindromic so as not to provide alternative secondary structure nucleation foci close to the test sequence. Of the commercially available restriction enzymes with non-palindromic recognition sites, there is none that does not cleave λ DNA. *Bsa*I was chosen as it only cuts the λ genome in two places, well away from the palindrome, providing fragments of recognizable sizes. The inserts all had the following form:

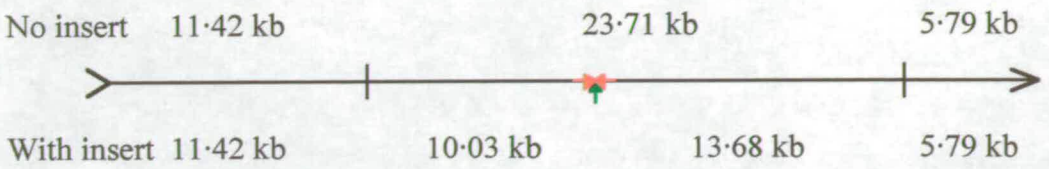


where $d(CAG)_n \cdot d(CTG)_n$ represents the test sequence. The *Bsa*I site was placed in opposite orientations on either side of the centre so that the palindromic sequence is continued right up to the test sequence. The 'phage were constructed as described in Chapter 2. Briefly, λDRL167 DNA was cleaved with *Sac*I, the insert was ligated in, the ligase was denatured and the DNA was redigested with *Sac*I to cleave 'phage without inserts before packaging the DNA and plating, selecting and plaque-purifying 'phage isolates. DNA minipreps of isolates were digested with *Sac*I and with *Bsa*I. Cleavage with these two enzymes produces the following fragments:

DRL167 with *Sac*I



DRL167 with *Bsa*I



The red arrows represent the palindrome and the green arrow the diagnostic cleavage at its centre. Representative gel sections are shown on the right. In isolate 1 insertion has not occurred but in isolate 2 insertion has been successful. There is some uncut DNA in both lanes 1 and not all of the 23·71 kb fragment of isolate 2 has cleaved, but the result is clear. When minigels were used, the 21·46 kb and 19·46 kb fragments of the *Sac*I digest were not resolved from one-another but were clearly further down the gel than the uncleaved genome; the *Bsa*I fragments were all resolved.

Results

'Phage were constructed as described and for the results given in this chapter plaque assays were conducted as in the 'early protocol' (Chapter 3). During this project many thousands of plaques were measured and it would not be practical to print all of the raw data but Figure 4.1 is included to convey an impression of the data from measurement of the plaques of a single 'phage. The figure (overleaf) shows a histogram of the areas of all the plaques measured and a cumulative frequency curve. Because of the spread of plaque sizes for any 'phage, cumulative frequency curves provided a much clearer means of showing the results of several sets of measurements on the same chart.

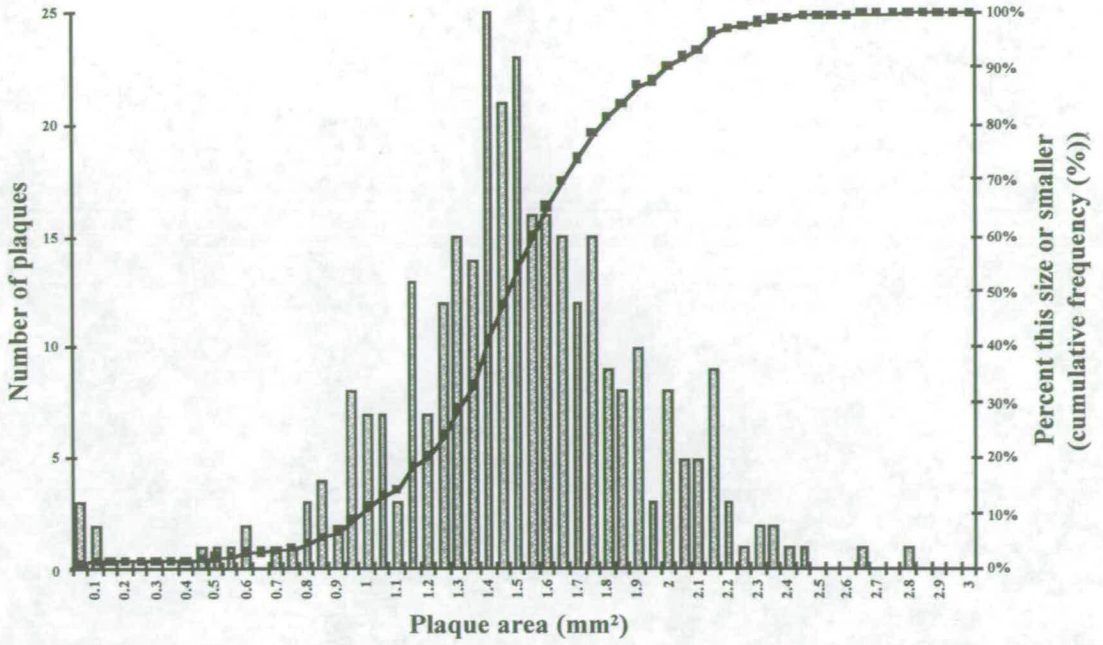


Figure 4.1 Area measurements of 319 plaques of 'phage 8,15 [$(GAC)_3$ centre].

For the very first constructions, oligonucleotides were just ordered to make inserts containing one d(CAG)·d(CTG) or one d(GAC)·d(GTC) trinucleotide in the centre and plaque areas were measured of two isolates of each of the 'phage constructed from them along with the 'phage, DRL176 which was included in each assay as a plaque size reference. Subsequently 'phage with two and three of each of these trinucleotides in the palindrome centre were constructed and another assay was conducted, and then the same was done with four and five copies of each trinucleotide.

Figure 4.2 (overleaf) shows cumulative frequency curves from these three plaque assays. It can be seen that in the first assay (a) plaque sizes were about the same for 'phage with a single copy of either trinucleotide in the palindrome centre and that the results for the two isolates of each 'phage were very similar to one-another. (As it happened the plaques were smaller than those of the reference 'phage, DRL176. It has a perfect palindrome with the central sequence d(GGATCC) ·d(GGATCC) but the sequence flanking that is not the same as the sequence flanking the trinucleotides under test so it is not directly comparable. It was only used as a

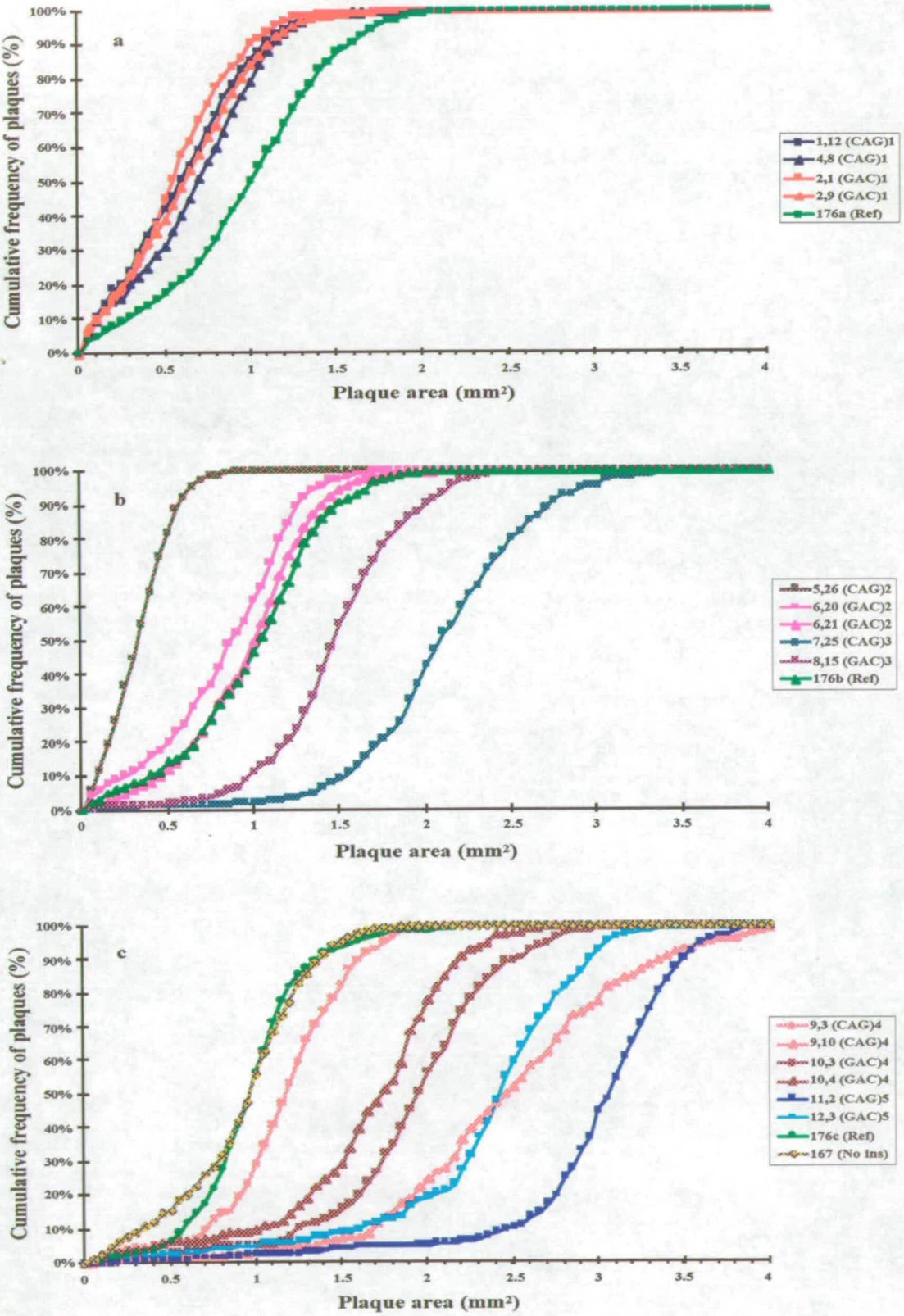
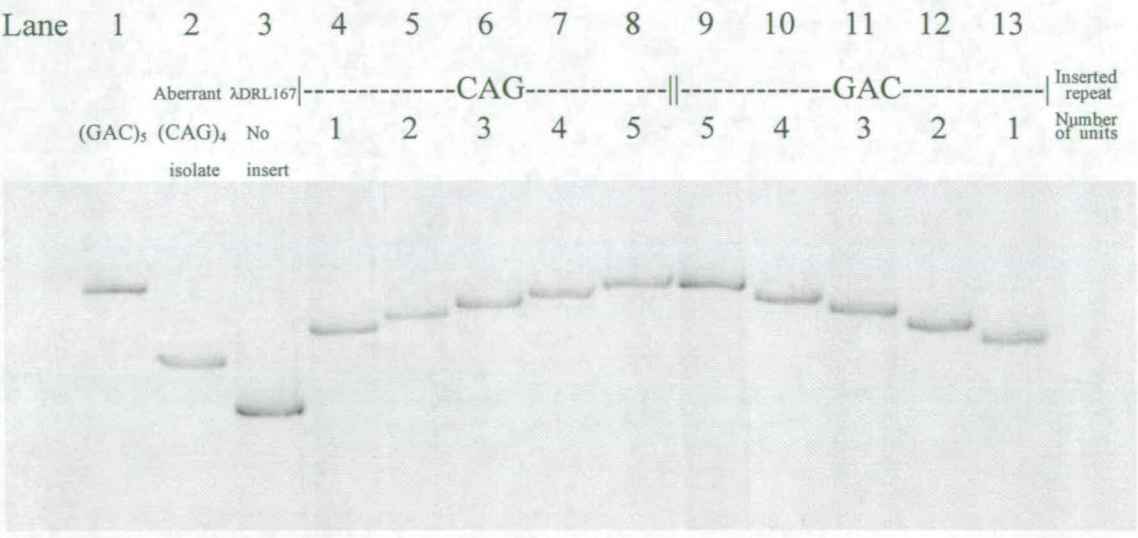


Figure 4.2 Cumulative frequency curves from the first three plaque area assays.

size reference for comparing different assays.) In the second assay (b) two isolates containing d[(GAC)·(GTC)]$_2$ were measured and again the results were very close. In the third assay (c) the parent 'phage DRL167 was included, just for interest, and its median plaque size (the area having exactly 50% of plaques larger and smaller than it) was seen to be the same as that of the reference 'phage. Plaques of two isolates containing d[(GAC)·(GTC)]$_4$ were measured. Their results were not as close as for the other pairs of isolates but the plaque size was larger and the difference in proportion to the size of the plaques was actually slightly smaller. However, the plaques of two isolates supposed to contain d[(CAG)·(CTG)]$_4$ were very different in size, 9,3 having a median plaque area of 1·17 mm^2 and 9,10 a median of 2·45 mm^2. It was clearly necessary to investigate. The palindromes were excised from the DNA of 'phage with all of the varieties of insert by the *Eco*RI sites at their ends, end-labelled and compared by polyacrylamide gel electrophoresis as described in Chapter 2. The PhosphorImage below shows the result.



Lane 3 contains the palindrome of DRL167 with no insert and lanes 4 -13 contain palindromes with inserts containing 1 - 5 d(CAG)·d(CTG) repeats and 5 - 1 d(GAC)·d(GTC) repeats. The length of the insert is (18 + 3N) bp where N is the number of central trinucleotides. Lane 1 is the same as Lane 9, just as a size marker. The d[(CAG)·(CTG)]$_4$ isolate whose palindrome appears in Lane 7 is 9,3. Lane 2

contains the palindrome of 9,10. The size shows that a deletion equal to approximately half of the insert has occurred. Restriction digestion had shown that at least one *Bsa*I site was present. Presumably the deletion was asymmetrical. This would account for the much larger median plaque size. (Having confirmed that the other constructs had inserts of the correct sizes, they were allotted the laboratory numbers DRL220 - DRL224 for $d[(CAG)\cdot(CTG)]_{1\text{-}5}$ and DRL225 - DRL229 for $d[(GAC)\cdot(GTC)]_{1\text{-}5}$ respectively.)

The median plaque area of the reference 'phage, DRL176 was almost the same in the three assays, 1.00, 1.04 and 0·97 mm^2 in a, b, and c respectively with 478, 337 and 341 plaques measured, so the other results from the three assays could safely be compared. Figure 4.3 (overleaf) shows the cumulative frequency curves for the series of d(CAG)·d(CTG) and d(GAC)·d(GTC) repeats separately. In (a) it is seen that even numbers of d(CAG)·d(CTG) repeats produce smaller plaques than do odd numbers and that the sequence $d[(CAG)\cdot(CTG)]_2$ gives very small plaques. Figure 4.3b shows that by contrast, the d(GAC)·d(GTC) repeats show little evidence of odd-even alternation. A continuous increase in plaque size is observed as the number of these repeats is raised from one to five. When median plaque area is plotted against number of repeat units, these patterns are demonstrated very clearly (Figure 4.4) except that the slight deviation from linearity with d(GAC)·d(GTC) repeats is actually seen better in Figure 4.3b.

The observation that central insertions of $d[(CAG)\cdot(CTG)]_2$ - and $d[(CGG)\cdot(CCG)]_2$ (see Chapter 6) - were responsible for the formation of very small plaques suggested that these sequences might be able to form tight loops. It was therefore decided to compare the effects on plaque size of a $d[(CAG)\cdot(CTG)]_2$ central sequence with central sequences that had been shown to form two- and four-base loops *in vitro* (Hilbers *et al.*, 1994) and *in vivo* (Davison & Leach, 1994b). For this comparison, two of the central insertions used by Davison and Leach (1994b) were reconstructed in the same sequence context used in the study presented here and plaques sizes

compared with the 'phage previously studied as well as with 'phage with a central $d[(CAG)\cdot(CTG)]_2$ or $d[(GAC)\cdot(GTC)]_2$ insert.

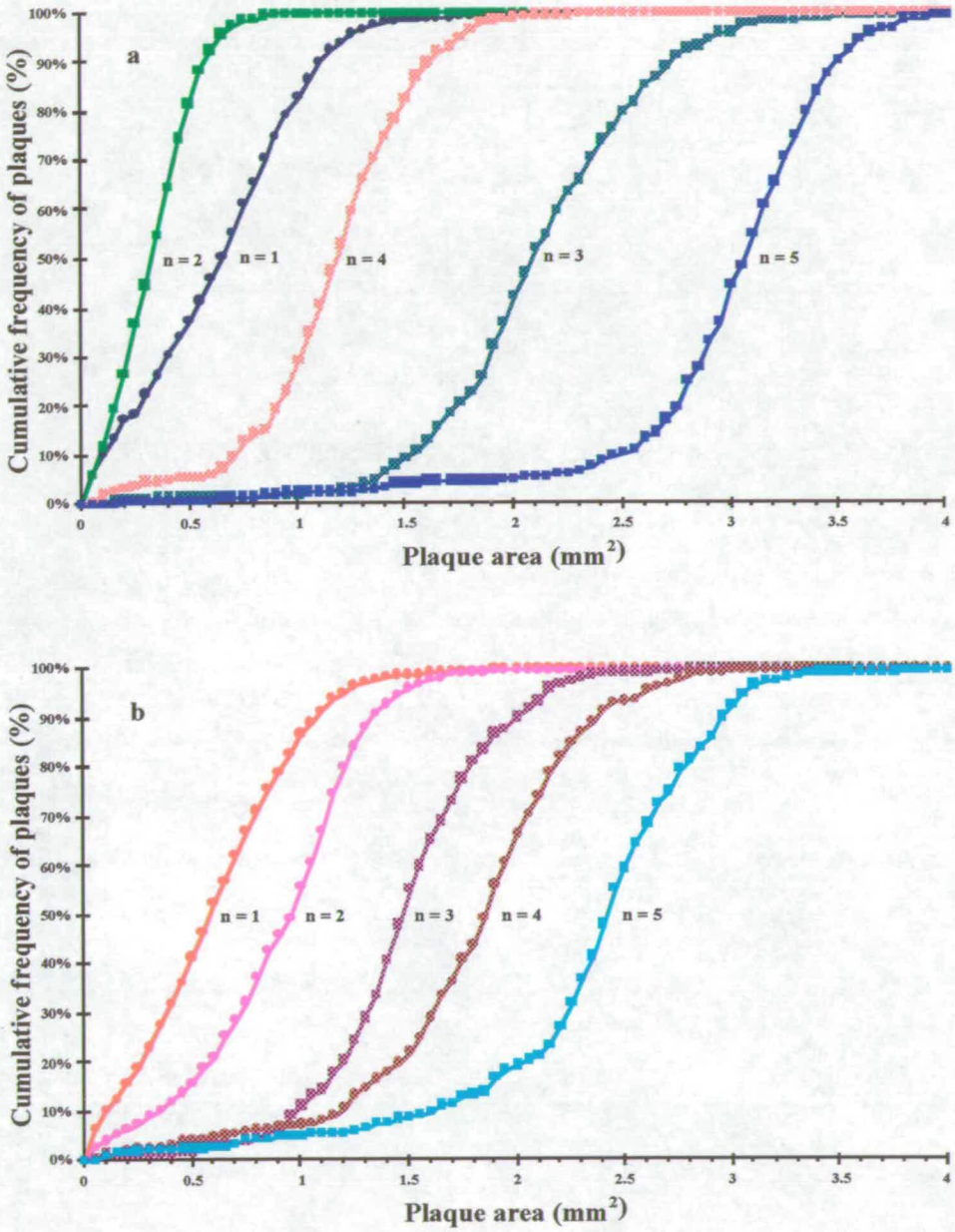


Figure 4.3 Cumulative frequency curves for plaque areas of (a) $d[(CAG)\cdot(CTG)]_n$- and (b) $d[(GAC)\cdot(GTC)]_n$-containing 'phage. The same colour-coding is used as in Figure 4.2 but data for isolates of the same 'phage have been combined (with the exception of the aberrant 9,10).

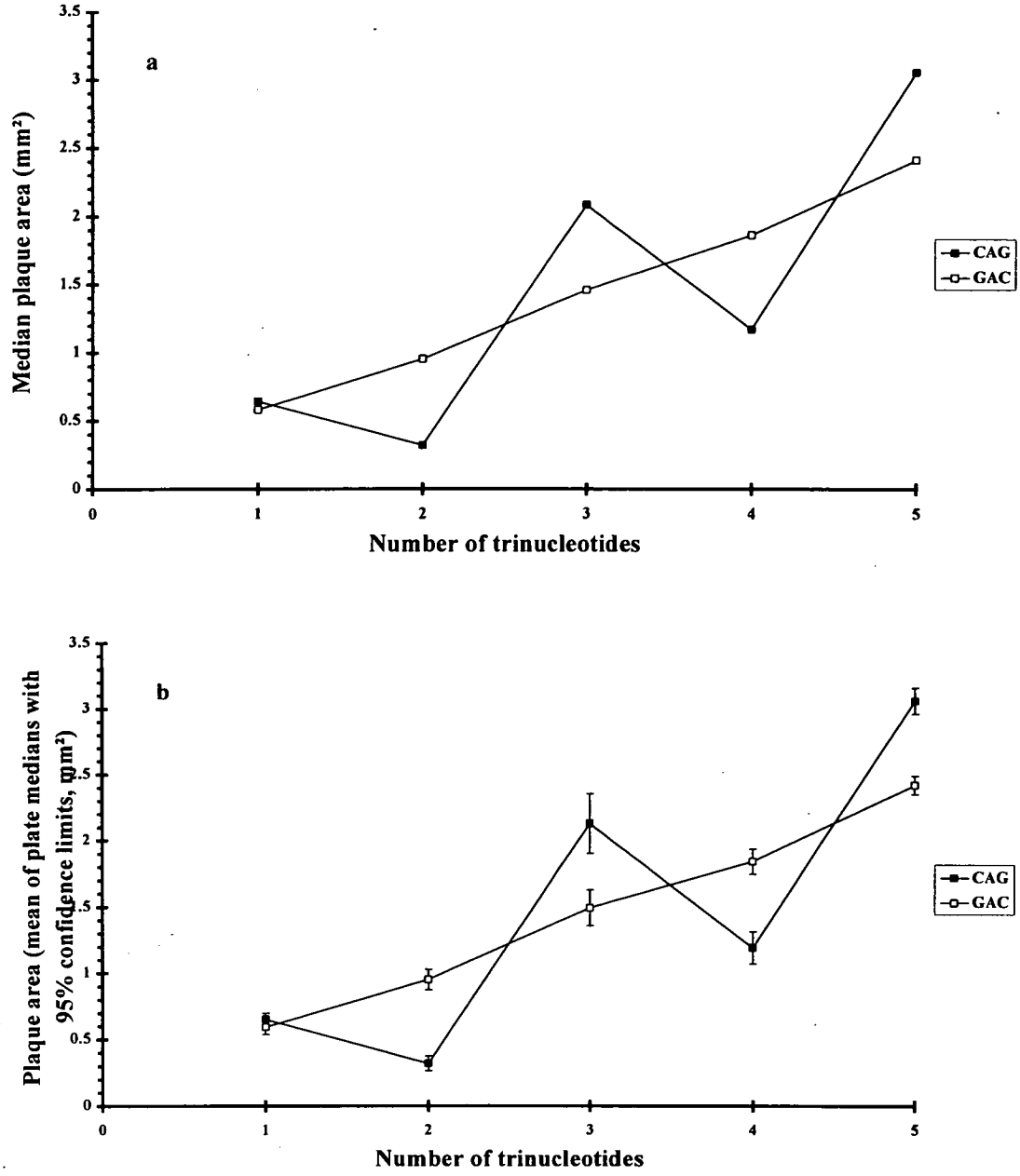


Figure 4.4 Median plaque area plotted against number of $d[(CAG)\cdot(CTG)]_n$ or $d[(GAC)\cdot(GTC)]_n$ trinucleotides: (a) the data presented as published (Darlow & Leach, 1995). After refining the plaque assay (Chapter 3) new assays were not performed on these 'phage but the data have been reprocessed as described in Chapter 3 and (b) shows the data presented as means of plate medians with 95% confidence limits. There is little difference in the shapes of the plots but some impression of the precision is given.

For all of the central inserts studied, it was necessary to make a choice for the two base-pairs flanking the $d[(CXG)\cdot(CX'G)]_n$ sequence. They were chosen to generate the sequence $d[G(CXG)_nC]\cdot d[G(CX'G)_nC]$ because a 5' G and 3' C would be present in a long array flanking a trinucleotide of this sequence. However, the sequence $d[(GAC)\cdot(GTC)]_n$ has a reversal of the G and C bases at the 5' and 3' ends of the repeat and a flanking 5' G and 3' C would not be the bases found adjacent to the trinucleotide in a repeated array. It was therefore considered possible that the small plaque phenotype conferred by $d(CAG)_2\cdot d(CTG)_2$ might be due to the nature of the flanking bases. Therefore 'phage were also constructed with $d[(CAG)\cdot(CTG)]_2$ and $d[(GAC)\cdot(GTC)]_2$ central sequences with the flanking bases the other way round. These were all compared in the same assay. The central sequences of the eight 'phage compared are listed in Table 4.1.

Central sequence	Source	Lab. No.
TGGAACTTGTTCCA	Davison & Leach, 1994b	DRL199
GGTCTCCACTTGTGGAGACC	This work	DRL235
TGGAAGTTCTTCCA	Davison & Leach, 1994b	DRL207
GGTCTCCAGTTCTGGAGACC	This work	DRL236
GGTCTCCCAGCAGGGAGACC	This work	DRL221
GGTCTCGCAGCAGCGAGACC	This work	DRL237
GGTCTCCGACGACGGAGACC	This work	DRL226
GGTCTCGGACGACCGAGACC	This work	DRL238

Table 4.1 Central palindrome sequences of one strand of the 8 'phage whose plaque sizes are compared in Figure 4.5. Outside the sequences quoted the palindromes are identical. The central sequence CTTG had been shown to form a two-base loop and GTTC to form a four-base loop.

The cumulative frequency curves are shown in Figure 4.5. It can be seen that the orientation of the bases immediately flanking $d(CAG)_2$ or $d(GAC)_2$ does not

greatly influence the plaque size. Also, the plaque size of 'phage with central CTTG or GTTC sequences is little altered by being put into the same context as the other sequences studied in this chapter compared with its previous context. The

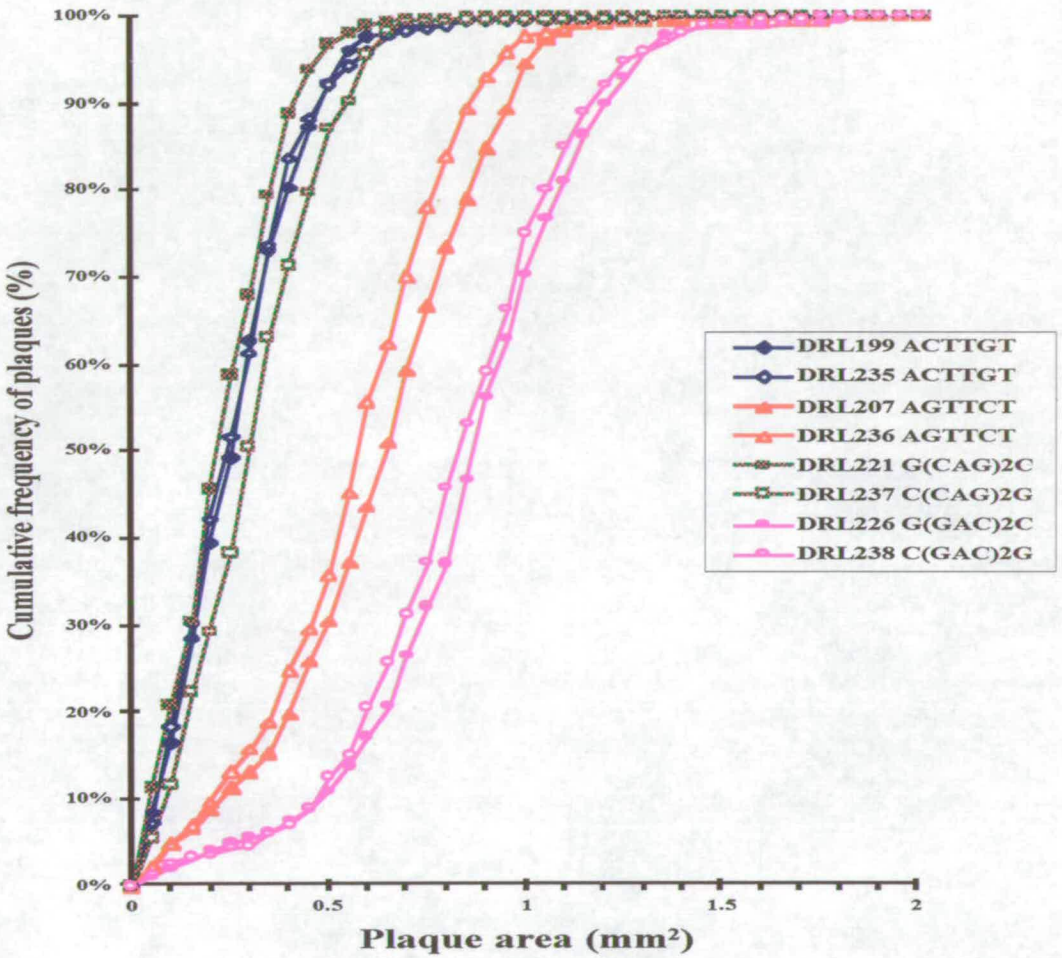


Figure 4.5 Cumulative frequency curves of plaque areas of 'phage with central sequences known to form a two-base loop *in vitro* (DRL199) and a four-base loop *in vitro* (DRL207) and six other phage for comparison (discussed in text). All the assays on this graph were performed at the same time but at a different time to the previous assays, hence some difference in the median sizes of the DRL221 [$(CAG)_2$] and DRL 226 [$(GAC)_2$] here from those in Figures 4.2, 4.3 and 4.4.

---

d[(CAG)·(CTG)]$_2$ central sequence confers a plaque size that is consistent with the formation of a two-base loop whereas the d[(GAC)·(GTC)]$_2$ central sequence gives

larger plaques even than those of the phage containing the sequence known to form a four-base loop, suggesting that it might prefer to form a loop with six unpaired bases.

Discussion

Analysis of the work described here

The plaque assay used here depends upon the finding that the plaque size in palindrome-bearing 'phage is acutely sensitive to changes in the central sequence of the palindrome. This suggested that a process similar to 'S-type cruciform extrusion' occurs *in vivo* (Davison & Leach, 1994a). In 'S-type cruciform extrusion' (first described in Lilley (1985) where it is referred to as 'Pathway B' and 'Mechanism B') DNA melting at the centre of the palindrome is followed by formation of a small 'proto-cruciform' and then branch migration results in the involvement of the whole of the palindromic sequence to make a larger cruciform structure if the protocruciform is stable enough (see Figure 4.6). The finding that in all positions outside the central

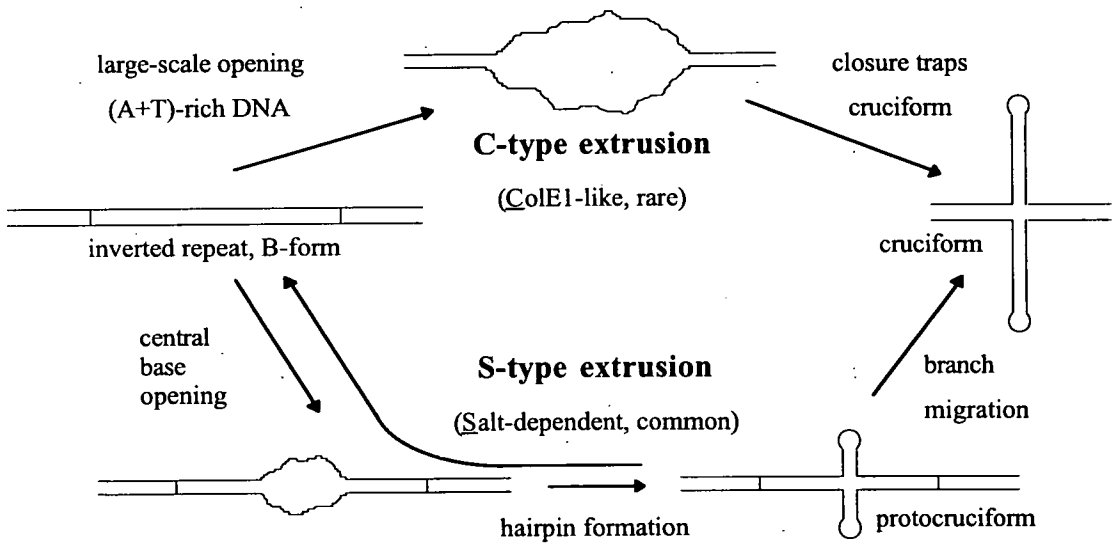


Figure 4.6 Alternative pathways for cruciform extrusion (after Murchie *et al.*, 1992)

two base-pairs of a palindrome C and G produced smaller plaques than A and T suggested that it is the stability of the protocruciform rather than the tendency to

central melting that is the more important in cruciform extrusion *in vivo* (Davison & Leach, 1994a). Thus the assay measures the relative abilities of central palindromic sequences to form stable hairpin-like structures.

Hairpin loop sequences of exceptional thermodynamic stability provide nucleation sites for folding of RNA (Varani, 1995) and the experiments of Davison & Leach (1994b, discussed below) with the plaque assay suggest that the same is true of DNA. The stability of a nucleic acid hairpin depends not only upon the length and the sequence of the stem, particularly the loop-closing base-pair, but upon the stability of the loop. The loop stability in turn depends not only upon its length, but upon its sequence, as these factors determine the possibilities of stacking of the loop bases with those of the stem and with each other, interactions of the loop bases with backbone sugar and phosphate groups, and the formation of non-Watson-Crick base-pairs (Hilbers *et al.*, 1994; Varani, 1995).

In general, shorter loops are more stable but this is not always evident from the sequence alone. For instance, in the series d(ATCCTA-T_n-TAGGAT), the T_4 loop was found to be more stable than the T_2 loop. However, further investigation showed that the A·T pair closing the -TT- loop was very unstable, if not totally disrupted, in which case there would really be a four-base loop, and in the sequence with the -TTTT- loop a T·T wobble base-pair was formed between the first and fourth T, making it effectively a two-base loop (Blommers *et al.*, 1987). Substitution of complementary bases for the first and last of the four thymines in this same context showed that base-pair formation is possible for YTTR hairpins but not for RTTY (where R indicates purine and Y pyrimidine), *i.e.*, 5' CTTG 3' and 5' TTTA 3' formed two-base loops but 5' GTTC 3' and 5' ATTT 3' remained as four-base loops (Blommers *et al.*, 1989). The authors were surprised to discover, however, that the T·A pair in -TTTA- was not a Watson-Crick base-pair but a Hoogsteen one; the C·G pair was Watson-Crick but quite distorted. Replacement of the two central thymines by the more bulky adenines limited the hairpin to a four-base loop (Blommers *et al.*, 1989). These investigations were done by NMR on single-stranded oligonucleotides.

In the plaque assay, both complementary strands are present. Davison & Leach (1994b) tested the central sequences d(CTTG)·d(CAAG), d(TTTA)·d(TAAA), d(ATTT)·d(AAAT) and d(GTTC)·d(GAAC) in the palindromic context given in Table 4.1 and found that the plaque size increased in the order just given but with a much larger difference between the YXXR and the RXXY than between the two YXXR or the two RXXY. This both indicated that the plaque assay agreed fairly well with *in vitro* findings and suggested that the plaque size depended upon the strand that formed the tighter loop. Agreement was not perfect however because the plaques of the 'phage with the central sequence d(ATTT)·d(AAAT) were slightly smaller than those of a 'phage with d(TTTT)·d(AAAA), though those for d(GTTC)·d(GAAC) were larger. Construction of 'phage with other bases for the XX above and outside the central four (Davison & Leach, 1994b; Davison, 1994) confirmed the general principle that YXXR sequences form more stable loops than RXXY sequences and suggested that the formation of two-residue loops *in vivo* may be more resistant to base sequence changes than *in vitro*.

Coming to the work of this chapter, $d(CXG)_n \cdot d(CX'G)_n$ repeats have the potential to form quasi-hairpins stabilized by C·G base-pairing and it was suggested that they can adopt either of two forms (Leach, 1994. See diagram on p. 71 of this thesis). One folds between d(CXG) units and contains an even number of repeat units and the other folds with an apical trinucleotide and has an odd number of units. If these structures are prone to form, it was predicted that one structure would be more stable than the other and that this would be determined primarily by the stability of the loop at the apex of the hairpin. Furthermore, if $d[(CXG)\cdot(CX'G)]_n$ sequences are particularly prone to form an unusual secondary structure, $d[(GXC)\cdot(GX'C)]_n$ sequences, which are not known to be prone to dynamic mutation, might not favour secondary structure formation. The work of this chapter was to compare the *in vivo* folding tendencies of odd and even numbers of d(CAG)·d(CTG) and d(GAC)·d(GTC) repeats.

The results revealed that even repeat numbers of d(CAG)·d(CTG) at the centre of a long palindrome produce smaller plaques than do odd numbers. This suggested that a favoured position of folding may exist between pairs of these trinucleotides to generate an even-membered hairpin-loop. The two types of loop are illustrated again in Figure 4.7. The strand containing thymines is shown because a strand containing pyrimidine-pyrimidine mismatches would be expected to form more stable loops and $d(CTG)_n$ was already known to form more stable hairpins than $d(CAG)_n$ (Yu *et al.*, 1995a,b; Gacy *et al.*, 1995). The figure is drawn to indicate the first three potential intra-strand base-pairs. The sequence $d[(CAG)\cdot(CTG)]_2$ gives very small plaques which suggested that it might fold into an unusually stable hairpin-loop. The sequence $d[(CGG)\cdot(CCG)]_2$ was also found to give very small plaques (see Chapter 6).

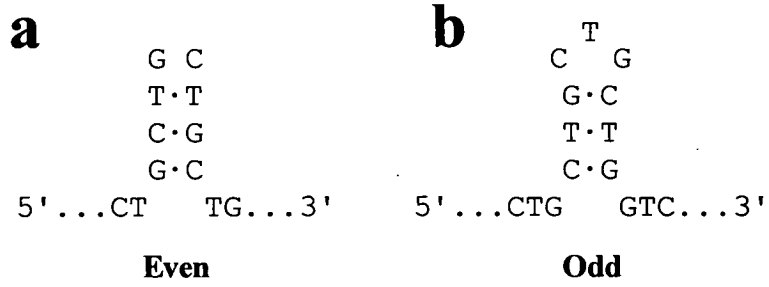


Figure 4.7 Alternative loops that may be formed by one strand of $d(CTG)_n$ in hairpins stabilised by C·G base-pairing.

These loops consist of a four-base sequence closed by a C·G base-pair (Figure 4.7a). The C is on the 5' side of the loop and the G on the 3' side, the arrangement of 5'-pyrimidine 3'-purine previously found to be favourable (Blommers *et al.*, 1989; Davison & Leach, 1994b), but here enclosing four bases rather than two. We felt (Darlow & Leach, 1995) that the thymines were likely to be accommodated easily within the helix to form both stacking interactions with the loop-closing base-pair and hydrogen-bonding interactions with each other as occurs in the wobble base-pair observed between thymines 1 and 4 in the d(TTTT) loop (Hilbers *et al.*, 1994). The structure might therefore share characteristics with two-base loops and is drawn with

a T·T wobble base-pair between bases 1 and 4 of the central loop. In order to determine whether the plaque-size data were consistent with the formation of a two-base loop, 'phage with central inserts of $d[(CAG)\cdot(CTG)]_2$ were compared with 'phage containing the central inserts d(CTTG) and d(GTTC) previously considered to form two and four-base loops (Davison & Leach 1994b). The results shown in Figure 4.5 are consistent with the formation of a loop containing two unpaired bases.

In Figure 4.7b an odd number of d(CTG) repeats generates an axis of folding which bisects the central trinucleotide. If a three-base loop forms, it will be closed by a 5' G·C 3' base-pair. Loops are intrinsic to RNA structure and they have probably been studied more in RNA than in DNA. Three families of unusually stable loop sequences occur over and over again in RNA, r(UNCG), r(GNRA) and r(CUUG) (Varani, 1995). Woese (1990) found that r(UUCG) is nearly always closed by 5' C·G '3 and r(GCAA) is usually closed by 5' A·U '3, but r(CUUG) is almost always closed by a 5' G·C 3' base-pair, and this work was confirmed and extended by Wolters (1992). One cannot assume, therefore, that because 5' G·C 3' was found not to be favoured for closing a loop of d(TT) (and various other two-base sequences) that it would not be favoured for a loop of d(CTG). The median plaque areas are so much larger for the 'phage with odd numbers of repeats than for those with even numbers that one might be tempted to speculate that the 5' G·C 3' base-pair does not form and that there is a seven-base loop d(TGCTGCT) closed by a 5' C·G '3 base-pair, but there is another possible explanation for the large difference.

With any number more than two d(CAG)·d(CTG) trinucleotides there will be competing $d(CTG)_2$ [and $d(CAG)_2$] pairs on either side of the centre, which may explain why $d[(CAG)\cdot(CTG)]_4$ plaques are larger than those of $d[(CAG)\cdot(CTG)]_2$. For odd numbers there are eccentric $d[(CAG)\cdot(CTG)]_2$ sites but no central site. Thus the plaques of 'phage with odd numbers of repeat units may be increased above the size of plaques that would result if there were no hairpin formation (by the repeat sequence) by a strong tendency for formation of even-membered hairpins at off-centre positions which would bring the palindrome arms together out of line and

probably result in the collapse of the protocruciform back to the normal duplex state. The trend is for increasing plaque size with increasing numbers of repeats but this does not necessarily suggest that longer hairpins are less stable, merely that with increasing numbers of repeats there are increasing numbers of eccentric positions for folding. It is for this reason, however, that the plaque assay is probably not suitable for testing the stability of structures that might be formed by long repeat tracts. Its application is in detecting the nucleating structures that could initiate the formation of larger structures.

No odd-even alternation of plaque size was observed for d(GAC)·d(GTC) repeats but there is the same the trend of increasing size of plaques with increasing numbers of repeats. This suggests that either there is no tendency of these repeats to form hairpins of 2 - 5 units or there is some tendency but odd and even loops are of almost equal stability and neither is stable enough to promote palindrome extrusion. From this one might guess that hairpins would be unlikely to form from a heteroduplex of this sequence. Whether this relates to the observation that this sequence is not found in expanded arrays has yet to be determined. The result does not contradict evidence that d(GAC) and d(GTC) can form hairpins when single-stranded but it does show that the introduction of more of these trinucleotides between the inverted repeats that are the palindrome arms renders cruciform extrusion steadily less likely. The results shown in Figure 4.5 suggest that both $d(GAC)_2$ and $d(GTC)_2$ remain unpaired as six-base loops. The 5' G and 3' C of these sequences are in the opposite polarity to the arrangement found favourable for closing the $d[(CAG)\cdot d(CTG)]_2$ loops. However, recent evidence discussed below suggests that there may only be four unpaired bases.

Further *in vitro* structural studies by NMR and melting

Since this work was published (Darlow & Leach, 1995) numerous other papers of varying degrees of relevance to this study have been published concerning repeats. Zheng *et al.* (1996) extended their earlier UV and NMR work (Smith *et al.*,

1995, discussed above) to cover all four d(CXG) repeats and both d(GXC) repeats but, as before, their oligonucleotides were very short, most work being done with sequences of three or four repeat units and NMR study of hairpin loops was not made. All six sequences were found to form antiparallel duplexes with right-handed helices. The d(GAC) repeats were found to form A·A wobble base-pairs with a single hydrogen bond under neutral to alkaline conditions but the adenines of d(CAG) repeats appeared to be conformationally unstable and only stabilized by protonation at low pH (5·4). d(GTC) repeats were found to have T·T base-pairs similar to those that they had reported earlier for d(CTG) repeats (Smith *et al.*, 1995). The combination of NMR work and UV melting studies suggested that both the d(GTC) and d(GAC) repeats form stably base-paired homoduplexes similar to the heteroduplex but with distortion of the backbones. In contrast all the d(CXG) repeat homoduplexes seemed to have smoother backbone conformations related to dynamic motions of the mismatched bases.

UV melting profiles also suggested that d(CXG) repeats were in hairpin ↔ duplex equilibrium from 4 repeats upwards and had an increased tendency to hairpin formation over the d(GXC) repeats. The stability of those with X = A or T was in the order CTG>CAG>GAC>GTC. The stabilities of all six homoduplexes were found to be less than those of the corresponding heteroduplexes but the differences were less for d(CXG) repeats than for the d(GXC) repeats. Heteroduplex → homoduplex transition was thus concluded to be easier for d(CXG) repeats but it was felt that the formation of secondary structures by single strands coming out of a heteroduplex would probably not be spontaneous but might be stabilized by the binding of specific recognition proteins.

Mariappan *et al.* (1996a) used nuclear magnetic resonance and computer modelling to study the structures formed by oligonucleotides of $d(CTG)_5$ and $d(CTG)_6$ and variants of these, made to determine which resonances belonged to which bases. The lengths were specifically chosen with the aim of producing hairpins with long enough stems for stability but not so long as to make it impossible

to distinguish individual bases. At the high DNA concentrations used for NMR, both were predominantly present as hairpins rather than duplexes so it was possible to study hairpins. Both of the oligonucleotides formed hairpins with no overhanging bases so they exemplified the two different types of loop. The studies confirmed the work of Smith *et al.* (1995) showing that T·T pairs form in the stem with two hydrogen bonds. Four possible configurations for the odd loop were investigated by computer modelling: with all three bases of the central trinucleotide on the 3' side, with one on the 5' and two on the 3', with two on the 5' and one on the 3', and all on the 5' side. In only one of these, that with all the bases stacked on the 5' side, was the 5' G·C 3' base-pair disrupted. It was not possible to distinguish the possibilities with certainty but the one which appeared to fit the NMR data best was that with all three bases stacked on the 3' side. The $d(CTG)_6$ hairpin was much more flexible than the $d(CTG)_5$ one. T·T pairing in the loop appeared to be present from the computer model but no evidence could be found for this in the NMR data and it was thought that the base-pair might be opening and closing so fast that the imino-proton signal could not be detected. Likewise the cytosine in the loop appeared to be very mobile. However, the study did not establish whether a hairpin with an even-membered loop is more or less stable than one with an odd-membered loop.

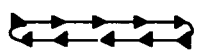
More recently the same authors have published, with others, their investigations of secondary structure of $d(CAG)_n$ oligonucleotides (Mariappan *et al.*, 1998a). They found that though hairpins were the major conformers of $d(CAG)_{5 \text{ and } 6}$ on gel electrophoresis, under the solution conditions required for NMR studies they were predominantly present as homoduplexes. The resonance overlap prevented high resolution determination of structure. $d(CAG)_{10 \text{ and } 11}$ were present exclusively as hairpins under NMR conditions but they too were too long for high resolution structure determination. Native polyacrylamide gel electrophoresis showed $d(CAG)_5$ present as both hairpin and duplex but $d(CAG)_6$ present only as hairpin so it appeared that within the range of 20 - 500 mM NaCl the $d(CAG)_6$ hairpin is thermodynamically more stable than the $d(CAG)_5$ hairpin. However, the authors did

not try a $d(CAG)_7$ oligonucleotide to see whether its hairpin was more or less stable than that of the $d(CAG)_6$ hairpin so the finding did not indicate whether the increased stability of the $d(CAG)_6$ hairpin over the $d(CAG)_5$ was due to its loop or its length, and the authors did not discuss the matter.

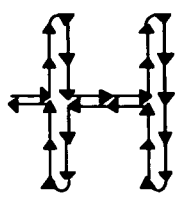
NMR of a $[d(CAG)_5]_2$ duplex with the adenines labelled at N6 with ^{15}N showed that the adenines were not extrahelical but could not determine whether they were hydrogen-bonded because of resonance overlap so a $[d(CGCAGCG)]_2$ duplex was examined. This contained the sequence $[d(GCAGC)]_2$ which occurs in duplexes of $d(CAG)_n$ and the results indicated that there was a single hydrogen bond between the mispaired adenine residues and the possibility of two hydrogen bonds was ruled out. It was therefore expected that the same would be the case in the stem of a $d(CAG)_n$ hairpin. Investigations of the non-exchangeable proton interactions in both duplexes were all consistent with B-DNA structure and strongly supported similar A·A geometries. Computer modelling was used to explore other possibilities for the arrangement of the adenines within the stem - a bifurcated hydrogen bond and no bond - to see whether they would fit as well or better with the NMR data but the data were most consistent with a single bond. This base-pairing seemed to be a consequence of C-A and A-G stacking rather than a driving force.

The authors mentioned their work in preparation for publication on $[d(GC\underline{GAC}GC)]_2$. This also showed a single hydrogen-bonded A·A pair also stabilized by stacking. The A-C stacking in this sequence was found to be stronger than the C-A stacking in the $d(CAG)_n$ duplex and the $[d(GC\underline{GAC}GC)]_2$ duplex was said to have normal B-DNA geometry for all the nucleotides. Both T·T and A·A bonds had already been examined by NMR at position N in d(GCCACNAGCTC)·d(GAGCTNGTGGC) by Gervais *et al.* (1995) and had likewise been found to have two and one hydrogen bond(s) respectively, both with only small changes to the normal B-DNA conformation.

Mariappan *et al.* (1998a) suggested that a $d(CAG)_{10}$ oligonucleotide might have one of three possible structures consistent with its fast electrophoretic mobility:

a hairpin with a four-base loop, a hairpin with a six-base loop and a 'dumb-bell' with two three base loops () where each arrow represents a trinucleotide, effectively two five-repeat hairpins end-to-end. The first would have one G residue in the loop and nine in C·G base-pairs in the stem, of which one would be susceptible to fraying and the dumb-bell would have two guanines in loops and eight in C·G pairs of which apparently none would fray. P1 digestion of $d(CAG)_{10}$ and $d(CAG)_{11}$ resulted in fragments which they estimated at 22 - 24, 14 - 16 and 7 - 8 for the former and 23 - 25, 15 - 17 and 8 - 9 for the latter. The longest fragments would not be expected from a single hairpin but would, they said, be expected from the dumb-bell, presumably from incomplete digestion cleaving only one loop.

Temperature- and pH-dependent ^1H NMR spectroscopy of $d(CAG)_{10}$ indicated two loop guanines and eight C·G base-pairs. This was clearly consistent with a dumb-bell. The authors said that two loop guanines would also be consistent with a single hairpin with a six-base loop but that in that case there would only be seven C·G base-pairs in the stem, but they miscounted. There would, of course, be eight. One might fray, but they did not say that. However, the P1 cleavage obviously did suggest a dumb-bell. It does not seem to have occurred to the authors, however, that if an even-membered loop is more stable than an odd-membered one, a dumb-bell could form that had four trinucleotides at one end and six at the other (there would be two isomers). The bands on their gel are so broad and fuzzy that I believe they do not exclude this possibility. The authors then suggested that longer $d(CAG)_n$ strands might fold into a tree-like structure:



(The authors drew all the individual bases.) Every branch in their diagram had five d(CAG) trinucleotides in it. The possibility of even loops was not mentioned.

Finally, Mariappan *et al.* (1998a) tried *in vitro* replication of $d(CAG)_8$ and of $d(CAG)_{21}$ in M13 single-stranded templates to see whether the nascent strand would show deletions due to folding of the template. Unfortunately they did their assay by cycle sequencing with *Taq* polymerase at 60 - 72°C. Replication of the template with $d(CAG)_8$ showed no deletion. Replication of the $d(CAG)_{21}$ template at 60°C showed a deletion of 4 repeats. The authors did not comment on the fact that this is an *even* number. Apparently they had a figure showing that this little hairpin showed progressive fraying as the temperature was increased, but the figure was left out of the paper. The authors pointed out that in their assay they had needed 21 repeats to demonstrate a deletion though oligonucleotides of 10 and 11 repeats exclusively formed hairpins. Amazingly, they concluded that this showed that the presence of the complementary strand and the polymerase pushed the critical threshold for hairpin formation to a higher value. They neglected to observe that human beings do not live in hydrothermal vents and that had they been able to perform their assay at 37°C they might have discovered larger deletions.

With reference to their conclusion that longer d(CAG) repeat single strands make multiple small hairpins rather than a single large one, Mariappan *et al.* (1998a) mentioned the work of Petruska *et al.* (1996). The latter examined the folding of 10- and 30-repeat oligonucleotides by UV melting profiles. The study was much more detailed than the melting studies of other groups (Yu *et al.*, 1995a,b; Zheng *et al.*, 1996) and included variation in DNA concentration, Na^+ concentration and rate of change of temperature as well difference between results obtained from DNA melted from a concentrated frozen stock and results after premelting or diluting.

The most interesting finding was that the stabilities of the secondary structures formed by $d(CTG)_{30}$ and $d(CAG)_{30}$ were little greater than those formed by only ten repeats of the same sequences. The melting temperatures being the same and the free energy changes being less than 40% higher in each case for the longer strands compared with the shorter ones. Thus at physiological Na^+ concentration (167 mM) the T_m of either $d(CTG)_n$ was 66°C and $\Delta G°$ (37°C) estimated at 4·7 and

6·5 kcal/mol for $d(CTG)_{10}$ and $d(CTG)_{30}$ respectively. Likewise the figures for the $d(CAG)_n$ strands were 60°C and 2·6 and 3·5 kcal/mol. The authors suggested that this might indicate that 30 repeats tend to form more complex hairpin folds with stems not much longer than those formed by 10 repeats and later suggested multiple short hairpins rather than a single long one. Since the slope of the melting curve is proportional to the number of base-pairs (2 refs quoted by Petruska *et al.*, 1996 earlier in the paper), and it is from the slope that the $\Delta G°$ is calculated, I imagine that the formation of two hairpins of 14 repeats each might fit the results but in the light of the findings of Mariappan *et al.* (1998a) perhaps the explanation is a dumb-bell with an average of 15 repeats in each end (*e.g.* 14 and 16).

Other findings of Petruska *et al.* (1996) were that the single strand structures formed by $d(CTG)_{10 \text{ and } 30}$ and $d(CAG)_{10 \text{ and } 30}$ and $d(GTC)_{10}$ and $d(GAC)_{10}$ were less stable than the respective complementary duplexes, in agreement with other work, and that increasing the sodium concentration from 19 mM to 167 mM increased the stabilities of all structures, as might be expected. Petruska *et al.* (1996) quoted references reporting that GC doublets, such as occur in CXG repeats have attractive base stacking interactions even in low salt whereas CG doublets, such as occur in GXC repeats, have less favourable stacking in low salt conditions. As salt concentration increases CG stacking becomes more attractive while GC stacking stays the same so that by ~1 M salt the two are equally stable. Petruska *et al.* (1996) comment that their results indicate that this difference in stability at the salt concentrations they used is maintained when the doublets are flanked by T·T mispairs but not when flanked by A·A mispairs. An interesting point that they did not mention is that CG doublets are more stable when flanked by A·A mispairs than when flanked by T·T mispairs. The order of stability of the ten-repeat structures was found to be CTG>GAC≈CAG>GTC (the GAC result slightly > CAG at both salt concentrations). The authors commented that this was a little different from the order derived from the electrophoretic mobility melting profiles of Yu *et al.* (1995a,b), with 15 repeats, which was GAC≈CTG>CAG≈GTC, the main difference

being that Petruska *et al.* (1996) found the T_m of $d(CAG)_{10}$ to be 47°C (at 19 mM), the same as for $d(CAG)_{30}$ whereas Yu *et al.* (1995b) found the T_m of $d(CAG)_{15}$ to be only 38°C. However, Yu *et al.* (1995b) had used only 1 mM Na^+. Petruska *et al.* (1996) had not seen the results of Zheng *et al.* (1996) which had come out presumably while they were in press. The latter derived the order CTG>CAG>GAC>GTC with their very short oligonucleotides.

The UV melting results of both Zheng *et al.* (1996) and Petruska *et al.* (1996) were complicated by the fact that in some curves a double sigmoid shape was seen, indicating melting of two different structures. Zheng *et al.* (1996) claimed to have found them with all four of the $d(CXG)_4$ oligonucleotides but not with the $d(GXC)_4$ ones. $d(CTG)_8$ showed only a single sigmoid shape and the authors said that the lower T_m of $d(CTG)_4$ and the T_m of $d(CTG)_8$ were constant through a 12-fold concentration range and it was from this that they concluded that the shorter oligonucleotide existed in hairpin ↔ duplex equilibrium. Thus it is clear that they believed the higher T_m of $d(CTG)_4$ to be that of a duplex. Petruska *et al.* (1996) found double sigmoid shapes with $d(CTG)_{10}$ and $d(GAC)_{10}$ but not with their other oligonucleotides. They found that the lower component appeared only when concentrated DNA suspensions were frozen and gradually disappeared as suspensions were diluted. It also became much less conspicuous with slower melting though this did not affect the major component. They then showed by electrophoresis of prewarmed and unwarmed samples of all their oligonucleotides that the minor components, *i.e.* the ones with the lower melting-points, were duplexes. Though hairpins were the thermodynamically preferred forms, Petruska *et al.* (1996) presumably could not bring themselves to believe that a side-by-side antiparallel duplex with more than twice as many base-pairs (and mismatches) as a hairpin could be less stable and so speculated that these unstable duplexes were formed from hairpins with one- or two- repeat overhangs that were loosely annealed end-to-end like restriction fragments. The reason for relating these findings will be seen later.

Structure revealed by use of DNA polymerases *in vitro*

1. Studies of slippage

Petruska *et al.* (1998) went on to make observations that directly support the finding of even-loop stability reported in this chapter but before describing this work, some introduction is needed. In November 1991, clearly unaware of the discoveries of the first two trinucleotide repeat disorders only a few months earlier, Schlötterer & Tautz (1992) submitted a paper which was nonetheless relevant. They were interested in the extensive length polymorphism of repeating sequences of 1-5-base units and cited slippage as the probable cause. They noted the early experiments in Khorana's laboratory (such as Kornberg *et al.*, 1964; Wells *et al.*, 1967a discussed in Chapter 1), showing that long repeat strands could be synthesized from two short primers by slippage, and set out to reinvestigate the phenomenon. They tried all ten possible trinucleotide repeats and two of the dinucleotide repeats, using in each case one primer of 15 nt and a complementary one of 9 nt, regardless of whether these were whole numbers of repeats, and a selection of polymerases all acting at 37°C. They incubated for up to two hours (with no cycling). The strands grew continuously until the reagents were used up. The products could be purified and used as templates for further synthesis and growth would start again at the original rate.

Trinucleotide repeats grew more slowly than dinucleotide repeats and all had different rates. d(AAT)·d(ATT) repeats grew most rapidly and d(GCC)·d(GGC) the most slowly, with d(GAC)·d(GTC) and d(CAG)·d(CTG) the next above. From this, along with examination of the sizes of the steps by polyacrylamide gel electrophoresis and an experiment in which one of the complementary strands was prevented from slipping at its ends by making it part of an M13 molecule, Schlötterer & Tautz (1992) conceived a model of slippage. They suggested that after melting of the 3' end of a strand from its template and reannealing further back, the little bulge behind it, which might be a single repeat unit looped out, could move as a wave in a 3'

→ 5' direction along the nascent strand to come off at the other end if not tethered. Furthermore, since the rate of extension was not length-dependent, the bulge did not have to reach the 5' end before another slippage could occur at the 3' end.

Since interest in trinucleotide repeats became widespread several studies of polymerase extension of primers on trinucleotide repeat templates have been published of which two were directly modelled on the study of Schlötterer & Tautz (1992) and were particularly interested in finding slippage of d(GCC)·d(GGC) repeats which had barely expanded in the earlier study. The first of these was by Behn-Krappa & Doerfler (1994) and they did try other sequences as well. They used PCR to extend a mixture of $d(CGG)_{17}$ and $d(GCC)_{17}$ and not surprisingly obtained expansion products. The authors gave no indication that they had realized that any repeating sequence would be bound to expand when subjected to cycles of melting, annealing, and extension (even if DNA were straight and rigid and completely incapable of the looping required for slippage) by the simple mechanism of annealing with overhangs and filling in. Then they tried PCR with each oligonucleotide separately and again obtained expansion products. They tried some other oligonucleotides alone and found that $d(CTG)_{17}$ and $d(CG)_{25}$ would expand and $d(TAA)_{17}$ and $d(CGGT)_{17}$ would not. They recognized that success depended upon the likelihood of annealing of identical sequences but did not appear to have considered that a single molecule might fold over and prime an extension on itself.

There was no indication that the authors had realized that though annealing in the first round of PCR had to be between two identical oligonucleotide sequences, the extension sequence would be complementary to its template, *i.e.* the extension of a $d(CAG)_n$ oligonucleotide would be with $d(CTG)_n$ and extension of a $d(CTG)_n$ oligonucleotide would be with a stretch of $d(CAG)_n$. Thus after several rounds of PCR all strands would have multiple alternating $d(CAG)_n$ and $d(CTG)_n$ stretches. The fact that alternating stretches of $(CAG)_n$ and $(CTG)_n$ or of $(CCG)_n$ and $(CGG)_n$ do not occur in human disease expansions indicates that polymerases do not tend to extend fold-backs *in vivo* (or that correction always occurs if they do).

Petruska *et al.* (1998) were inspired by this peculiar experiment. They said, tactfully, merely that the work of Behn-Krappa & Doerfler (1994) was the only previous study of self-priming with trinucleotide repeats. Petruska *et al.* (1998) set out deliberately to induce self-priming by hairpin formation (rather than by intermolecular annealing) by the two means of keeping the DNA concentration low and making the 3' end fully complementary to the template repeats near the 5' end. They used an oligonucleotide with the sequence $d(CTG)_{16}(CAG)_4$ in order to form a hairpin in which the top of the stem and the loop were entirely composed of d(CTG) repeats but the other end of the stem was held by perfect complementary base-pairing. They incubated for periods of up to an hour with a DNA polymerase, initially the proofreading-deficient Klenow fragment exo$^-$, at 37°C and ran the products on a denaturing polyacrylamide gel. If a molecule folded to form a blunt hairpin it would not be extended. If it folded leaving a 5' overhang this would be filled in by the polymerase and a longer molecule would result. If there was subsequent slippage, further extension would occur with time. If the molecule folded with no overhang or with an even number of trinucleotides overhanging there would be an even-membered hairpin-loop. If it folded with an odd number of trinucleotides overhanging there would be an odd-membered hairpin-loop.

The picture that emerged was that hairpins with even-membered loops vastly outnumbered ones with odd loops. The overhangs were filled in within seconds so that at 0·5 min there was a ladder of molecules with 0, 2, 4, 6, 8 and 10 trinucleotides added. Bands showing odd numbers of trinucleotides added were very faint or invisible. This supports the findings of this chapter (and the authors mentioned this). With increasing incubation times the proportions of longer products increased (and proportions of shorter ones correspondingly decreased). Thus the hairpin with no added sequence was at a peak at the start and declined thereafter. The one with two repeat units added reached a peak at ~1·5 min and then declined. The one with four added units reached a peak at ~4 min, and declined and so on.

Hairpins with more than 10 units added took much longer to appear, the band for 12 added units not reaching the intensity of 0·5 min bands of 2 - 10 added units until about ten minutes and the 14-added-unit band not for at least 20 min. The reason for this again lay partly in the nature of the loop. With 12 added d(CAG) repeats the molecule would have the sequence $d(CTG)_{16}(CAG)_{16}$, *i.e.* a d(CAG) repeat would have to move into the loop before the last two repeats could be added. Thus the apical six nucleotides forming the loop would have the sequence d(CTGCAG) which would probably not be as stable as d(CTGCTG) because of restricted space in the loop. The loop of the +14-unit molecule would have the sequence d(CAGCAG) and reached the intensity of the +10 of 0·5 min at somewhere between 20 and 40 min of incubation. At 40 min, the main bands were of +10, +12 and +14. The smallest was a very faint +9.

The +13 and +15 hairpins (the heaviest band seen at the top of the picture) would both have had an odd membered loop with d(CAG) in the centre, the former having d(CTG) on one side of it and d(CAG) on the other and the latter having a d(CAG) on both sides of it. Both of these bands were much weaker than the +14 band indicating that even-membered loops are more stable than odd-membered ones in $d(CAG)_n$ just as they are in $d(CTG)_n$. The authors hinted that they may be working currently to produce a corresponding set of results from a $d(CAG)_{16}(CTG)_4$ oligonucleotide. That seems almost superfluous, but what I really hope they will do is to try $d(GAC)_{16}(GTC)_4$. I say this rather than $d(GTC)_{16}(GAC)_4$ because the GAC strand makes the more stable hairpins so it is the possibility of any odd-even difference in loops made by this strand that would be required to check my result.

There was another reason for the slower appearance of higher-molecular-weight bands. From measurement of the intensities of the bands in the first ladder at 0·5 min and at subsequent times it was possible to calculate rate-constants for slippage from position $0 \rightarrow 2$, $2 \rightarrow 4$ *etc.* and this showed that the rate of slippage was in direct proportion to the number of d(CTG)·d(CTG) pairings and in inverse proportion to the number of d(CAG)·d(CTG) pairings. Thus a hairpin formed from

a single strand looped from complementary duplex DNA will slip more easily upon itself than upon its complementary template.

In the initial ladder of bands of DNA increasing in units of two repeats seen at 0·5 min one might have expected a continuous decline in abundance from 0 to 10 units indicating declining stability of hairpins with longer overhangs. In fact the +4-repeat band was the weakest (5% of the DNA) and Petruska *et al.* (1998) suggested that this might be because longer overhangs were stabilized by folding over to form another hairpin. This would fit with the dumb-bell idea and results of Mariappan *et al.* (1998a) (except that the loops are now shown to be predominantly even membered and not odd-membered as the latter imagined). The fact that in the initial population a substantial proportion of the hairpins had overhangs shows that the idea of their previous paper that duplexes might be formed by the annealing of overhangs of two hairpins was a possibility. The melting curve of $d(CTG)_{16}(CAG)_4$ was obtained and it too had a double sigmoid shape, the lower inflexion being at 56°C and the higher at 80°C. This time the explanation given was that the lower T_m represented melting of the d(CTG)·d(CTG) part of the stem and the upper one was for the d(CAG)·d(CTG) part on the ground that $d(CTG)_6(CAG)_4$ had only the higher T_m.

Petruska *et al.* (1998) repeated their polymerase experiment with the usual proof-reading-proficient Klenow fragment of DNA polymerase I and found that the 3'-exonuclease activity of proof-reading had a marked inhibitory effect upon expansion. With their previous dNTP concentration of 0·2 μM, the ladder at 1 min of incubation was similar except for two deletion bands but with time the expansion bands diminished in intensity and more deletion bands appeared. A five-fold increase in dNTP concentration to 1 μM allowed expansion to occur beyond the +10 repeats seen at 1 min, but a further ten-fold increase to 10 μM was needed to give the same results as for the deficient enzyme.

2. Searches for synthesis arrest in repeat tracts

In view of evidence of secondary structure formation and synthesis aberrations and arrest in repeat tracts, and delay in replication of expanded alleles in fragile-X syndrome, Wells and colleagues (Kang *et al.*, 1995) investigated the replication of d(CAG)·d(CTG) and d(CGG)·d(CCG) repeat tracts *in vitro*. They initially used a plasmid containing $d[(CAG)\cdot(CTG)]_{130}$ in its human genomic context from the *DMPK* gene, denatured with alkali, added a primer, neutralized to reanneal, then incubated for 10 min at 37 or 50°C before adding a polymerase and incubating for a further 10 min. They tried Klenow fragment, Sequenase (a modified form of T7 polymerase) and in one experiment human DNA polymerase β and found with all of them, on electrophoresis of the products, what they referred to as 'pause' sites meaning in fact that many products stopped but some went on. The different enzymes stopped in different positions.

The apparent blockages occurred whether the d(CAG) or the d(CTG) was the template. They were stronger after the lower preincubation temperature and weaker as the length of the repeat tract was diminished. They disappeared if preincubation was at 70°C, and if synthesis was carried out after preincubation at 37°C and then the DNA was incubated at 70°C followed by cooling and adding more polymerase, more longer products were produced. The blockages did not occur if a single-stranded template was used but did not require supercoiling because linearized plasmids gave the same results as intact ones. Use of 7-deaza-dGTP (which can make the same Watson-Crick bonds as dGTP but cannot make the hydrogen bonds involved in tetraplexes and triplexes) made no difference but use of dITP instead of dGTP caused all the stalling to occur at the beginning of the tract. Similar apparent blockages were seen with double-stranded d(GCC)·d(GGC) repeats but only if the d(GCC) strand was the template. Subsequently similar results were obtained with d(GAC)·d(GTC) repeats, only when the d(GTC) strand was the template, and for d(GAA)·d(TTC) and d(GGA)·d(TCC) repeats (Ohshima *et al.*, 1996a), but not for any of the other trinucleotide repeats (Ohshima *et al.*, 1996a,b)

The inference of all this was that there were secondary structures that formed in some double-stranded repeat DNA tracts which arrested the progress of replication, but there was one very strange finding: the further back was the primer from the beginning of the repeat tract, the further the polymerase seemed ~~manage~~ to get into the repeats and the distance run beyond the beginning of the repeats was around 20 bp greater than the distance from the primer to the beginning of the repeats (Ohshima *et al.*, 1996a). The puzzle was explained when stalled products were isolated and sequenced by the Maxam-Gilbert method (Ohshima & Wells, 1997). The results showed that the polymerase had proceeded a short distance into the repeats and then the new strand had folded over to pair with itself and the rest of the product was complementary to the 5' flanking sequence and ended at the 5' base of the primer.

With d(CAG)·d(CTG) repeats with the d(CTG) strand as the initial template, three major products from each of two primers were sequenced and for each primer the products contained 8, 10, and 12 d(CAG) repeats. The complementary flanking sequences showed that they had all formed hairpins in which the first repeat unit was paired with the last so that all had even-membered loops. With one of the primers a fourth major product was seen that was 6 bp longer than the one below so presumably had 14 repeats with another even loop. Chemical modification with bromoacetaldehyde (BAA) or diethylpyrocarbonate (DEPC) followed by cleavage with piperidine, to look for unpaired bases, revealed that the hairpin loops consisted of the four bases d(AGCA) closed by 5' C·G 3'. This is perfectly in keeping with my results. The loop formed by the d(CAG) strand was not expected to be as tight as that formed by its complement. The A residues in the stem were modified by DEPC but were less vulnerable towards the base of the stem than near the loop. Ohshima *et al.* (1997) took this to indicate that longer lengths of d(CAG) repeats probably form more stable hairpins. This idea does not fit with the results of Petruska *et al.* (1996) that suggested that stability does not increase with length beyond some point not far over 10 repeat units. The greater protection of the A residues near the base may just

relate to the fact that, counting the C residue of the first triplet and the G residue of the last, the $d(CAG)_n$ hairpins were closed by $C_6 \cdot G_6$ in the flanking sequence. Ohshima *et al.* (1997) did not isolate and sequence d(CTG)-containing strands.

With d(GAC)·d(GTC) repeats with the d(GTC) strand as the initial template, three major products (from just one primer) were sequenced. Unlike those from the d(CTG) template, these were only separated in size by 3 bp each suggesting approximately equal tendencies to form hairpins with odd- or even-membered loops. The authors counted the repeats as $d(CGA)_n$ but products were found to contain 7, 8, and 9 d(GAC) units paired d(GAC)·d(GAC) in the stem, two with odd-membered loops and one with an even-membered one. Chemical modification and cleavage revealed seven unprotected bases in the odd loops - d(ACGACGA) - and only four in the even loop - d(ACGA). Since both types of loop were closed by the same d(CG)·d(CG) doublet it is hard to see how they could be of approximately equal stability unless there is some kind of base-pairing within the odd-membered loop. My suggestion that is supported by the finding that though all the A residues in the loop were hypersensitive to DEPC, cleavage at the first C residue of the loop after BAA modification was less than that at the second by what looks to be well over an order of magnitude. There also seems to be somewhat less cleavage of the second G residue (that might pair with the first C) than of the first. Also, Yu *et al.* (1995b) found cleavage with P1 nuclease only at the GpA and ApT positions of the central trinucleotide and at the CpG 3' to it. One can see that the A residues in the stems were far better protected than those in the stems of the d(CAG)·d(CAG) hairpins which supports studies suggesting that d(GAC)·d(GAC) hairpins are more stable than d(CAG)·d(CAG) ones. The results with d(CGG)·d(CCG) repeats will be mentioned in the next chapter.

The question that remains is of why synthesis stalled in the repeat tracts, allowing the 3' ends of the nascent strands to fold over. Ohshima *et al.* (1997) only investigated this with d(CGG)·d(CCG) templates. Mytelka & Chamberlin (1996) investigated the sequences at which Sequenase tends to 'pause' in sequencing

reactions and the ways in which the problem might be ameliorated. They compared 21 nt sequences around pause sites, ranked in order of severity and found that from 15 severe pause sites they obtained a consensus containing a single d(YCG) trinucleotide (Y = pyrimidine) in a region that tended to be GC-rich, and that there was a tendency of pauses to occur near other pauses. A computer search for folding showed that the sites were often near possible hairpins but not in any consistent position in relationship to them, some being found at the beginnings of stems, some in loops and some at the ends of stems. Mytelka & Chamberlin (1996) found that betaine and some related compounds could relieve the pausing. Ohshima *et al.* (1997) tried incubating with 2M betaine before elongation with Klenow fragment and found no improvement with the d(CGG)·d(CCG) templates. Higher concentrations only caused general inhibition of synthesis. Single-strand binding-protein did reduce the stalling, but this does not tell us whether there were structures involving single-stranded DNA in the template; the protein might merely have prevented the folding of the nascent strand back onto itself. I will return to this problem in the following sections.

Flexibility of d(CTG)·d(CAG) repeat tracts

Gel mobility has been used to characterize three- and four-way junctions, quadruplex DNA, flexible DNA and especially bent DNA. In agarose, bent and 'straight' DNA migrate similarly but in polyacrylamide bent DNA migrates more slowly than expected (refs in Chastain *et al.*, 1995). Chastain *et al.* (1995) investigated the gel mobility of double-stranded d(CAG)·d(CTG) and d(CGG)·d(CCG) repeat tracts of different lengths, with and without interruptions, with flanking sequences, to look for evidence of anomalous helical structure. The investigations showed that the tracts migrated up to 20% more rapidly than marker DNA. Using a formula from a 'reptation' model (a hypothesis that DNA migrates in the manner of 'a snake in a burrow') this was consistent with a 20% increase in 'apparent persistence length' (P_{app} or P_a) derived from relative migration rate.

Persistence length is a measure invented in 1949 for describing the flexibility of chain molecules, involving projecting all the bonds of the chain onto the first bond and taking an average for all the various shapes of the molecule, and has been variously reformulated since for use with DNA (see Schellman, 1974; Frontali *et al.*, 1979 and refs therein and refs *there*in). The apparent persistence length is that experimentally observed and is partitioned by the equation $1/P_a = 1/P_s + 1/P_d$ where P_s, the static persistence length, is the mean length over which static axial deflections sum to an angle of one radian and P_d, the dynamic persistence length, is the mean length over which thermal fluctuations, in the absence of static bends, sum to an angle of one radian (Chastain & Sinden, 1998).

Chastain *et al.* (1995) reasoned that if the rapid mobility of d(CAG)·d(CTG) repeat DNA was related to its helical structure then disruption of the structure by the DNA intercalator actinomycin D might reduce the mobility (relative to marker DNA) and this proved to be the case. Relative to markers, the rate of migration increased with increasing acrylamide concentration and with decreasing temperature but was not simply related to the length of flanking DNA. In relation to the position of the repeat tract within flanking DNA, the mobility was greatest when the tract was in the middle rather than near one end of the molecule. Rate of migration increased with length of the repeat tract and this was not affected by interruptions. Since DNA molecules containing hyperflexible regions had been found to migrate more slowly the authors suggested that the increase in apparent persistence length might indicate decreased 'bendability' and concluded that the results indicated a special helical structure.

Bacolla *et al.* (1997) used plasmid restriction fragments containing ranges of numbers of d(CTG)·d(CAG) [and d(CCG)·d(CGG)] repeats to measure their tendencies to form circles relative to the rates of ligation of fragments prevented from forming circles by being made blunt at one end. This depends not only upon length but upon the bending modulus, the torsional modulus, the persistence length and the helical repeat (bp/turn) of the sequence. Their results indicated that both types of

repeat are around 5 - 20 times as flexible as random sequence DNA but suggested that they are nonetheless fully paired duplexes. The flexibility was reflected by a *low* persistence length and a low bending modulus while the torsional modulus and helical repeat were found to be close to normal. The efficiency of circularization was found to be ~1,000 times more than for random sequence DNA; the decrease in persistence length caused an approximately 40% drop in the optimal length of circularization (552 bp for random DNA to 326 bp for d(CTG)·d(CAG) and 366 bp for d(CCG)·d(CGG) repeats).

Calculations of 'writhe' (the out-of-plane trajectory of the helix axis in circular DNA), observations of the numbers of different topological isomers of repeat-containing plasmids due to different numbers of supercoils, and observations of the effect on the calculated number of base-pairs per turn from electrophoresis of random-DNA plasmids with different numbers of repeats inserted all indicated that the repeat DNA has greater variance of writhe than random DNA, $d[(CTG)\cdot(CAG)]_n$ > $d[(CCG)\cdot(CGG)]_n$, and tends to act as a sink for superhelical density. Attempts at cleavage of plasmids containing various lengths of the repeats after modification with seven different chemical agents as well as digestion with S1 nuclease and DNase I showed no evidence of accessible bases or unpaired regions. The authors (Bacolla *et al.*, 1997) also quoted five other papers from the same laboratory, plus some new data not shown, that electrophoresis had shown no evidence of supercoil-induced structural transitions, and another of their papers reporting that immunological tests were also consistent with there being no unusual secondary structures or single-stranded regions.

Bacolla *et al.* (1997) point out that trinucleotide repeat DNA is structurally repeated every two helical turns (3 bp × 7 = 21 bp and one turn ≈ 10·5 bp) so might be expected to be straight. Their values for apparent persistence length P_a, 278 Å for $d[(CTG)\cdot(CAG)]_n$ and 315 Å for $d[(CCG)\cdot(CGG)]_n$, are ~60% of P_a (~500 Å) or ~40% P_d for random-sequence DNA, and this suggested that one or more dinucleotide repeat steps within the sequence might be more flexible than average. They reviewed

X-ray crystallographic results on the ten possible dinucleotide pairs and noted that d(CA)·d(TG), d(CG)·d(CG), d(GC)·d(GC) and d(CC)·d(GG) came second to fifth in order of flexibility, d(AG)·d(CT) coming eighth. They then used all the crystallographic results to construct an order of flexibility of all ten possible trinucleotide repeat sequences plus the two mononucleotide repeats d(AAA)·d(TTT) and d(CCC)·d(GGG) in degrees of angle and found $d[(CCG)\cdot(CGG)]_n$ and $d[(CTG)\cdot(CAG)]_n$ to come third and fourth, with $d[(ACC)\cdot(GGT)]_n$ and $d[(AAC)\cdot(TTG)]_n$ first and second, $d[(GTC)\cdot(GAC)]_n$ fifth and $d[(GAA)\cdot(CCT)]_n$ eleventh. Thus flexibility clearly cannot be the most important factor in repeat expansion, though it might help. The repeats that came first and second in flexibility might not be expected to form hairpins. The $d(GGT)_n$ strand of the first would be expected to be able to form tetraplexes (see Chapter 5), which it has now been shown to do (Usdin, 1998), and this repeat was found to undergo large expansions (Lindblad *et al.*, 1994), though the repeat is not as common in genes (Stallings, 1994).

In an accompanying paper (Gellibolian *et al.*, 1997) some of the same authors present further calculations based upon the same results with circularization of DNA. They found that the differences in the variances of writhe of d(CTG)·d(CAG) and d(CCG)·d(CGG) repeats from that of random B-DNA are greatest at about 700 - 800 bp, *i.e.* about 230 - 270 repeats. The free energy of supercoiling (lower for d(CTG)·d(CAG) than for d(CCG)·d(CGG) repeats and both lower than for random sequence) turned out to have maximum differences from random DNA at similar lengths: about 167 repeats for d(CTG)·d(CAG) and about 180 for d(CCG)·d(CGG). This length - in which the DNA has the greatest tendency to writhe, they named the 'region of hyperflexibility' and pointed out that it corresponds approximately to the boundary between premutation and full mutation for DM and fragile-X syndrome, the point beyond which massive expansions occur.

In the meantime, Chastain & Sinden (1998) had second thoughts about their conclusion of increased stiffness of these repeat DNAs because the previously-demonstrated preferential assembly of d(CTG)·d(CAG) repeats into nucleosomes

would be consistent either with the DNA being curved or flexible and the DNA was unlikely to be curved because curved DNA migrates more slowly. They interspersed short d(CTG)·d(CAG) tracts between short sequences with known straight or curved helix trajectories to see how the mobilities of these sequences were altered by the trinucleotide repeats and tried interspersing known torsionally-flexible and 'bendable' DNA sequences between known curved DNA for comparison. From these experiments they concluded that the trinucleotide repeats are not intrinsically bent and behave as 'bendable' and torsionally flexible joints, four repeats being similar in effect to interspersion of the double mismatch d(TT)·d(TT). They also tried circularization of d(CTG)·d(CAG) repeat DNA and this agreed with the results discussed above that the DNA is highly flexible. They then tried to rationalize these conclusions in the face of other work demonstrating reduced electrophoretic mobility of flexible DNA and amongst these thoughts was the possibility that the previously-tested sequences (phased 1-3-base gaps and nicks) might become kinked whereas the d(CTG)·d(CAG) repeats would not.

S-DNA

The possibility of slippage of repeats in complementary duplex DNA to form loops of each strand in different places was discussed in Chapter 1. Pearson & Sinden (1996) set out to investigate this possibility by inserting trinucleotide repeat tracts into a plasmid, denaturing with alkali, then neutralizing and allowing the DNA to renature. With d(CTG)·d(CAG) they used three genomic clones from the DM locus including flanking sequence. Of these, one had 17 repeat units, one had 50, and one had about 251 repeats plus four d(ACT)·d(AGT) interruptions, representing normal, premutation and full mutation lengths respectively. In one experiment the plasmids were linearized by cleavage on one side of the repeats before denaturation and renaturation and then cleaved on the other side of the repeats before electrophoresis. In addition to bands the same as those seen with DNA not denatured (more rapidly moving than random DNA), all three repeat lengths gave rise

to more-slowly-moving smudges of DNA (with some major bands) that were subsequently (Pearson *et al.*, 1998a) shown to consist of many discrete bands.

In order to check that these products could result from reannealing of strands from the same original duplex the experiment was repeated without linearizing the plasmids before denaturation and renaturation. This gave rise to the same bands though they were much fainter. Further restriction digestions closer to the repeats showed that the anomalous migration did map to the repeats. The fact that the structure would form from linearized plasmids showed that superhelical tension was not required for their production.

These structures proved to be very stable. When the major anomalous bands were cut out of the gel, eluted, ethanol-precipitated and phenol-extracted with strong vortexing they still retained their same mobilities in comparison with unpurified DNA. They also withstood heating at 55°C for an hour, but after heating at 85°C for an hour each of the three major anomalous bands of 255 repeats gave rise to bands of the other mobilities and ones of the mobility of the undenatured DNA, and previously undenatured DNA gave rise to the anomalous bands. The DNA structures from the anomalous bands also withstood overnight incubation at 55°C and freezing at –20°C, and only after prolonged radiographic exposure was some interconversion (~2%) seen.

By differentially end-labelling the two strands taking part in these structures, the authors were able to show that the $d(CAG)_n$ strand was slightly more sensitive to mung-bean nuclease than the $d(CTG)_n$ strand. There was quite a high level of background cleavage of both strands, repeating or non-repeating, denatured and reannealed or not, but with the 'reduplexed' DNA of both the 50-repeat and the 255-repeat tracts, a single, rather faint extra band was seen in the $d(CAG)_n$ lanes after the strands had been separated and irreversibly denatured by glyoxal. Since labelling was at one end, this does not necessarily indicate that cleavage was in only one place. Mung-bean nuclease cleavage of the 17-repeat tract was not tried. Since these structures were formed from complementary strands of repeats bounded on both

sides by non-repetitive DNA the logical conclusion was that they were formed by misalignment of repeats with looped-out sections on each strand making hairpins, cruciforms or other structures and Pearson & Sinden (1996) named this S-DNA (for slipped-strand DNA). They pointed out that looped out sections might not only be in different places but be different in number and size on the two strands and might even form different structures.

Further investigation (Pearson *et al.*, 1998a) has shown that the proportion of the repeat DNA forming more-slowly-migrating configurations on denaturing and renaturing increases steeply from about 2% for 17 repeats to a plateau of about 70% at somewhere a little below 50 repeats. Because a small proportion of their plasmids contained deletions or expansions in the repeat tracts, Pearson *et al.* (1998a) cut out specific bands and subjected the DNA to a second round of denaturation and renaturation which showed that these structures did form in DNA in which the two strands were the same length (though one would have thought that the experiment with circular plasmids might have been sufficient). They then examined reannealed DNA by electron microscopy. DNA of the perfect 50-repeat sequence, from a major band that had a mobility of a tract 1·25 times as long, showed randomly positioned bends and kinks and length measurements showed that the molecules were shorter than undenatured DNA and in a range of roughly 3 - 31 repeats looped out (not necessarily all in the same place). In DNA from a band with a mobility of a tract 2·0 times as long, unusual secondary structures were visible. These included cruciforms, three-way junctions and apparent open loops and Θ-like structures, both thought to be formed by the interaction of two hairpins. Similar types of structures were seen in the 255-repeat DNA. The limit of resolution was for structures of about 15 repeats and some were estimated to contain up to 50 repeats.

Pearson *et al.* (1998a) referred to work that showed that denaturation and renaturation of d(GA)·d(TC) and d(CA)·d(TG) repeat DNAs resulted in formation of novel products but that these were formed by mispairing of multiple strands. They had tried formation of S-DNA with these sequences themselves and had not found it

to occur, suggesting that not all tandemly repeated sequences would form S-DNA. They announced that analysis of sequence dependence of S-DNA formation on various di- tri- and tetra-nucleotide repeats is in progress. Though theoretically it could form in any tandemly repeated DNA one might expect that it would be most likely when (a) the length of the repeating unit is short and (b) individual strands are self-complementary - which is not the case with the two dinucleotide repeats mentioned. In fact this work is not entirely new; the authors acknowledge EM work on misalignment in minisatellite DNA with repeating units of 17 - 37 bp going back to 1989, but the structures formed were not biophysically stable.

Though Wells and colleagues (Bacolla *et al.*, 1997) reported that they had never found any evidence of unusual secondary structure formation occurring in d(CTG)·d(CAG) or d(CCG)·d(CGG) repeats under superhelical stress, one cannot help feeling that, since the flexibility of these repeats predisposes them to accumulate supercoils, S-DNA might form in them *in vivo* under certain conditions - apart from those involving obligatory denaturation, including transcription. Kohwi *et al.* (1993) demonstrated that in the presence of Zn^{2+} or Co^{2+}, all the cytosine residues but none of the adenine residues of d(CAG)·d(CTG) repeats became vulnerable to chemical modification. This would occur at pH 7 and in the presence of Na^{+} and Mg^{2+} but not if neither Zn^{2+} nor Co^{2+} were present. Conceivably this might indicate that these repeats have a liability for loops to pop out.

The literature on trinucleotide repeats is vast, and growing daily. In this chapter I have limited discussion mainly to results relating to secondary structure formation rather than to the mechanism of expansion. Some of the many other papers, and hypotheses put forward in some of the above-mentioned papers, will be mentioned in the final chapter.

Late addition

Since proof-reading this chapter another paper of relevance has come out and I have decided to discuss it here rather than integrating it into other parts of the

chapter. Having come to the conclusion that the major instability in the disease-causing trinucleotide repeat tracts is not due to defective mismatch repair and must be due to some structure formed in the DNA (Goellner *et al.*, 1997), Gacy & McMurray (1998) set out to investigate the reason for the dramatic increase in instability ~~of~~ beyond a threshold length which they cite as 29 - 35 repeat units for Huntington disease. They say that, while slippage is more probable in long repeat tracts, it is only modestly increased in this range and that another mechanism must come into play. They present the puzzle that if this mechanism involves the formation of a large hairpin it has to be explained why a perfect palindrome of the same length is not equally unstable.

First, Gacy & McMurray (1998) carried out polyacrylamide gel electrophoresis and determined melting temperatures by absorbance at 260 nm, both in the presence of 100 mM NaCl at pH 7. They tried duplexes of 10 and 25 repeats of d(CAG)·d(CTG) and control sequences with the same bases but randomized in order within each trinucleotide, and 25 repeats of d(CGG)·d(CCG), d(GAC)·d(GTC), and the non-hairpin-forming sequences d(AAG)·d(CTT) and $d[(CA)\cdot(TG)]_{37}$, and concluded that the duplexes of trinucleotide repeat disorders do not display unique properties in these respects. Melting was approximately according to GC content and perhaps the percentage of their gels, distance run, and length of the repeat tracts were not enough to demonstrate the 20% greater mobility noted by Chastain *et al.* (1995), Chastain & Sinden (1998). They also found that hairpins of those sequences that would form hairpins varied in stability. Their order for 25 repeats of those sequences discussed in this chapter was GAC (54·9°C) > CTG (51·4°C) > CAG (50·1°C) > GTC (43·0°C) (*c.f.* results on pp. 164-165). They did not quote Petruska *et al.* (1996) but, like them, found that the melting points for 10 and 25 repeats were the same, both for d(CAG) and for d(CTG).

Gacy & McMurray (1998) then investigated whether hairpin formation by single strands reduces the rate of formation of complementary duplexes. Single complementary oligonucleotides were allowed to equilibrate in 100 mM NaCl, pH 7,

at 37°C for half an hour, and then mixed, and the rate of hairpin→duplex conversion was monitored both by absorbance at 260 nm and by electrophoresis. For the 75 bp sequences (*i.e.* 25 trinucleotides or 37 dinucleotides) the rates of conversion for those that form hairpins were 1 - 2 orders of magnitude lower than for those sequences that would not form hairpins whereas at 30 bp, the rate for 10 repeats of d(CAG) and d(CTG) was only slightly lower than for the pseudo-random sequence of the same bases. From these results, Gacy & McMurray (1998) concluded that short and long tracts of d(CNG) repeats form hairpins of equal stability (based upon T_m), and form them just as rapidly, but that longer tracts are liable to expansion because of the much longer lifetime of their hairpins in the presence of the complementary strands. Palindromes, they concluded, also form hairpins of long lifetime, but form them much less rapidly, and therefore much less often, because of the exact pairing required. (They put it that "An entire half of the palindrome must be unpaired before hairpin formation is possible." That of course is not so, but the hairpin or cruciform can only nucleate by melting exactly at the centre whereas nucleation of a quasi-proto-hairpin of repeats can occur at numerous places in the tract. This however is not the reason why palindromes do not often lead to expansion, but I shall leave discussion of that to the final chapter.)

Not quoting Petruska *et al.* (1996), Gacy & McMurray (1998) completely avoided the question of how hairpins of 10 and 25 repeats could have the same melting point and yet the one reanneal to its complement much more rapidly than the other. They did note that there were only two bands in each lane of the gel monitoring the hairpin→duplex conversion, showing no detectable presence of intermediate products. Presumably because of their discovery of the retarded reannealing, they did not consider the possibility that the equal melting points might mean that longer oligonucleotides form multiple short hairpins. However, perhaps the branched structures suggested by Mariappan *et al.* (1998a) might provide an answer. Perhaps the situation might be analogous to the problem of trying to handle a long piece of sticky tape. If for a moment one relaxes ones guard and allows two

parts of the tape to come in contact, and finds that one has to release ones fingers from the ends of the tape to pull apart the local self-adhesion, whilst one is attending to this, other parts of the tape may stick together. Thus with the longer single strands of trinucleotide repeats, while raising the temperature to the melting point will result in all the branches of the structure melting simultaneously, melting of any part of the structure at 37°C may just lead to reannealing into the same or a slightly different shape. It has to be said that the hairpin band on the gel is a discrete band and not a smudge so does not allow of a range of different structures with different mobilities, but 25 repeats is not very many and the branched options would be similar and not numerous. Gacy & McMurray (1998) did not say why, they used oligonucleotides of only 25 repeats when they had stated the threshold for expansion in larger jumps to come at 29 - 35 repeats, but I note that the oligonucleotides that they used were mainly ones that they already had from their work published in 1995.

Like Petruska *et al.* (1996), Gacy & McMurray (1998) calculated free energy and enthalpy changes for the melting of their secondary structures. Taking into account the difference in Na^+ concentrations of the buffers, the two groups had similar results for the 10-unit oligonucleotides but the latter had rather higher estimates for the longer oligonucleotides, giving $\Delta G°_{37}$ 81% higher for $d(CTG)_{25}$ than for $d(CTG)_{10}$ and 140% higher for the respective $d(CAG)_n$, but they did not comment on that. Gacy & McMurray (1998) did however state, as nobody has ever done before to my knowledge, that $d(CTG)_n$ and $d(CAG)_n$ form hairpins of similar stability, though their own $\Delta G°_{37}$ figures actually show nearly twice as much free energy for the CTG strand than for the CAG strand at 10 repeats and about 47% more at 25 repeats.

Other results, experiments and ideas from this paper that relate to secondary structure of the single strands of d(CGG)·d(CCG) repeats and to the mechanism of expansion of trinucleotide repeat tracts will be discussed in the next and the final chapter respectively.

Chapter 5

Review of *in vitro* work on secondary structures in d(CGG) and d(CCG) repeat tracts

Introduction

While reading the relevant literature in preparation for publishing the results of Chapter 6, on hairpin folding of d(CGG)·d(CCG) repeats *in vivo*, I found that there was so much disagreement amongst investigators of the secondary structures formed by the single strands of this sequence *in vitro* that a short discussion in a paper would not give enough room to explore the reasons for these disagreements so as to arrive at a best guess at what might happen under truly physiological conditions *in vivo*. It was therefore decided to write a separate review (Darlow & Leach, 1998a) to be submitted with the paper (Darlow & Leach, 1998b). This chapter is essentially the review with some modifications, corrections and additions, including discussion of papers that have come out since.

One problem was that most authors considered only a subset of all the possibilities of hairpin folding of these strands and there was no agreed nomenclature. The full range of possible arrangements was therefore laid out and named as shown in Figure 5.1 (overleaf) and this scheme will be used in the following discussions. Each possible alignment is defined in terms of the frame in which the sequence 5'→3' is the same on both sides of the hairpin.

In alignments with frames 1 and 2, two out of three bases in each trinucleotide are involved in Watson-Crick base-pairs and the remaining base is in a C·C or G·G mispair depending upon the strand. Frame 1 [d(CGG)·d(CGG) and d(CCG)·d(CCG)] [= alignment (a) of Mitas *et al.* (1995b), Yu *et al.* (1997b) but alignment B of Gao *et al.* (1995)] is akin to the pairing $d[(CAG)\cdot(CAG)]_n$ and $d[(CTG)\cdot(CTG)]_n$. Frame 2 [d(GGC)·d(GGC) and d(GCC)·d(GCC)] [= alignment

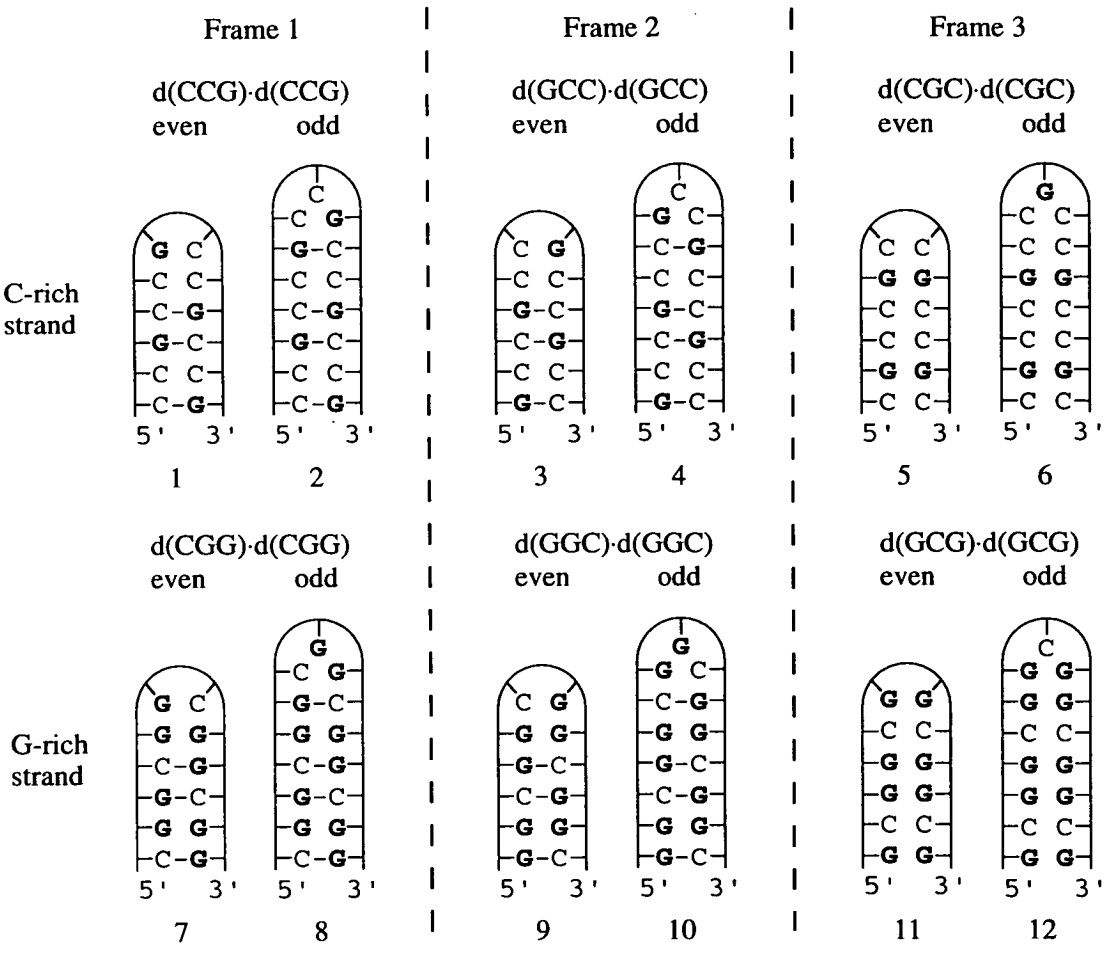


Figure 5.1 All the possible types of hairpin loops that might be formed by single strands of d(CGG)·d(CCG) repeats. Only Watson-Crick bonds are shown but other workers have found evidence of G·G bonds and C·C bonds *in vitro* (see text). N.B. The above classification is based upon the alignment of the two sides of the hairpin and the presence of an odd or even number of unpaired bases in the loop. It does not depend upon the actual number of unpaired bases in the loop, length of the stem or the 5'-base of the sequence. The alignment is defined by the frame in which the sequence 5'→3' is the same on both sides of the stem. For all three alignments some workers have postulated that long hairpins might fold over to form unistrand quadruplexes.

(b) of Mitas *et al.* (1995b), Yu *et al.* (1997b) but alignment A of Gao *et al.* (1995)] is akin to the pairing $d(GAC)_n \cdot d(GAC)_n$ and $d(GTC)_n \cdot d(GTC)_n$. In Frame 3 [d(GCG)·d(GCG) and d(CGC)·d(CGC)] [= alignment C of Gao *et al.* (1995)] there

are no Watson-Crick base pairs. In the C-rich strand this alignment is intuitively unlikely. However, attention has been drawn to this alignment because of the possibility that the G-rich strand might be able to pair this way and form a quadruplex structure held together by G_4-'quartets' (Figure 5.4c). There has been much interest in the potential for formation of such structures by G-rich sequences that occur in telomeres (Venczel & Sen, 1993; Williamson, 1994; Kettani *et al.*, 1995). It has been shown that such structures can form *in vitro* from four DNA strands, from two hairpins, and from a single strand. It has not been proven that they occur *in vivo* with telomeres but there is evidence of quadruplex formation *in vivo* with a G-rich sequence upstream of the human insulin gene (Hammond-Kosack *et al.*, 1992a,b,c).

These quadruplexes require the presence of cations for their formation. In the case of univalent ions, a single cation sits between two G_4-quartets in an octahedral complex with the carbonyl groups of the guanines. The divalent cations have been thought to promote DNA structures with tight helices by acting as counterions between the phosphate oxygens of adjacent backbones (refs in Venczel & Sen, 1993) but Venczel & Sen (1993), noting the similarities of the stabilizing orders univalent and divalent ions, raised the possibility that the divalent ions might also be complexed between the guanine quartets. The divalent cations achieve their effect with about two orders of magnitude lower concentrations than the univalent cations (Venczel & Sen, 1993; Lee, 1990) but the best effect achieved by any ion is dependent upon the van der Waals radius of the hydrated ion, not just its charge and concentration, and K^+ fits best into the octahedral cage. The Mg^{2+} ion is only about the same size as that of the Li^+ ion. It has a larger effect upon stability because of its higher charge but still has a lesser effect than Na^+. The Ca^{2+} ion is similar in size to that of Na^+. Thus the order of stabilizing ability for the main intracellular cations is $K^+ > Ca^{2+} > Na^+ > Mg^{2+}$ (Hardin *et al.*, 1992).

For d(CGG) repeats it has been suggested that hairpins might form in frame 3, held together by G·G Hoogsteen bonds and possibly by C·C bonds in addition. It

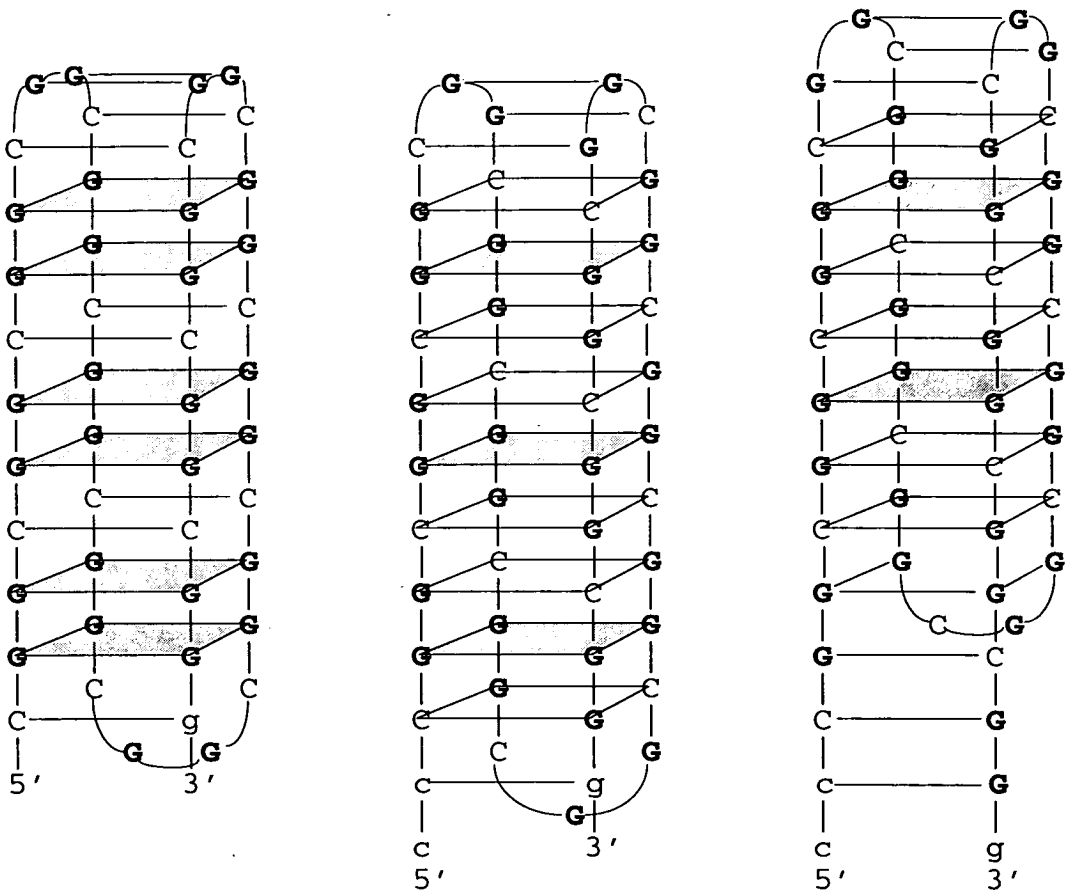


Figure 5.2 Examples of quadruplexes in the three alignments of pairing. (b) and (c) are the two quadruplexes suggested by Mitas *et al.* (1995b) for $d(CGG)_{15}$ redrawn to give an impression of three dimensions. The bases indicated by small letters are flanking bases of their construct. (a) is a quadruplex in Frame 3 that might be formed by the same sequence, $d(CGG)_{15}$ (and one of the flanking bases). The front left strands of all three structures are aligned. G_4-quartets are shaded to emphasize the different patterns and C·G·C·G· quartets are indicated by unshaded rhomboids.

---

has also been suggested that long hairpins of this type might fold over onto themselves to form quadruplexes. A diagram of such a structure is shown in Figure 5.2a. Mitas *et al.* (1995b) have suggested that hairpins of the G-rich strand in frame

1 or 2 might similarly fold to form quadruplexes and that these might contain C·G·C·G· quartets. In either frame there would be two C·G·C·G· quartets to every one G_4-quartet. Diagrams of examples of these quadruplexes are given in Figure 5.2, b and c. It has been shown that quadruplexes containing C·G·C·G· and G_4-quartets can form *in vitro* in the presence of Na^+ (Kettani *et al.*, 1995; Kettani *et al.*, 1998) but it is not yet established whether they can exist in the presence of the larger K^+ ion.

This chapter examines all of the data now available on secondary structure in the single strands of d(CGG)·d(CCG) repeats to seek to discern the natures of the most stable structures formed by each strand. The discussion divided roughly into sections on each of the two strands but the division cannot be complete as inevitably some work compares the two. DNA structure is affected by type and concentration of cations, by pH, and by temperature. It will be seen that different investigators have worked with very different values of all of these so it is worth remembering that what are relevant to repeat expansion causing human disorders are the conditions inside the nucleus of a human cell. pH can be measured by intracellular electrodes but, because these have to pierce the cellular and nuclear membranes, ions may leak. Therefore more reliable estimates are obtained by using indicator dyes. Estimates of the intranuclear pH vary from 6·0 to 7·3 though in most cells it is towards the upper end of this range. The principal cation in the cell is potassium and $[K^+] \approx 150$ mM. $[Na^+] = 10 - 30$ mM depending upon the activity of the cell. The ratio of magnesium to potassium is 1:5·67, giving the former a total concentration of about 26.5 mM but most of this is bound, the free $[Mg^{2+}]$ being about one tenth of the total. Likewise most calcium is bound; the free ion concentration can vary by three orders of magnitude between $\sim 10^{-8}$ and $\sim 10^{-5}$ M. (Information from Dr. Malcolm O. Wright, Department of Physiology, University of Edinburgh, personal communication.) Temperature, of course, is around 37°C in the ovaries and usually a few degrees lower in the testes.

The G-rich strand: frame 1, 2 or 3?

Investigations giving evidence for alignment in frame 1 or 2

Mitas *et al.* (1995b) considered two alignments, frame 1 [= '(a)'] and frame 2 [= '(b)']. They used an oligonucleotide (excised from a plasmid) containing the sequence $dCC(CGG)_{15}G$. They considered the possibilities of hairpins 8 and 9, both with and without G·G bonds, and quadruplexes formed by either of the two hairpins folded over on itself and bonded by one G quartet and two C·G·C·G· quartets per repeat (Figure 5.2, b and c). By methylation of the self-annealed oligonucleotide with DMS (dimethyl sulphate) followed by cleavage at the modified bases and electrophoresis under denaturing conditions it was possible to investigate the nature of the secondary structure. DMS methylates the N7 positions of guanine residues. In G·C bonds, and in the C·G·C·G· quartet arrangement of Mitas *et al.*, the N7 positions are not involved in hydrogen bonding; in G·G bonds the N7 position of one of the guanines is involved, and in G_4-quartets the N7s of all four residues are hydrogen-bonded and thereby protected from methylation. No guanine residues were completely protected under a wide range of conditions which was deemed to rule out quadruplex structure. Residues in the stem of a hairpin or quadruplex are protected relatively to those in unpaired loops. A hairpin has one loop of unpaired bases whereas a unimolecular quadruplex has three (Figures 5.2 and 5.5). DMS and P1 nuclease, which also attacks these loops, both indicated that there was only one loop, again suggesting hairpin structure. Relative reactivities of bases in the loop decreased with increasing KCl, indicating that potassium stabilizes the structure.

A melting study showed that the T_m of $d(CGG)_{15}$ is 27°C higher than that of $d(CTG)_{15}$ from which the authors (Mitas *et al.*, 1995b) concluded that G·G base-pairs contribute a significant amount of stability to the hairpin. Relative methylation of the two G residues of the GpG dinucleotide can distinguish between the alignments because in frame 1 it is the 5' Gs that are involved in G·G bonds whereas

in frame 2 it is the 3' ones. This study indicated that at KCl concentrations of ≥ 200 mM the hairpins were all aligned in frame 2 (hairpin 9) but that at concentrations of ≤ 100 mM this hairpin must be in equilibrium with another structure which offers less methylation protection to the 3' G residues. This could either be the same hairpin but without G·G bonds or it could be a hairpin 8 structure with G·G bonds.

An NMR study (Chen *et al.*, 1995; Mariappan *et al.*, 1996b) concluded that pairing of short $d(GGC)_n$ oligonucleotides was in frame 2 under all conditions tested and that the mismatched G residues were strongly paired and stacked with the neighbouring G·C pairs. The authors were unable to examine loop structure by NMR because of the predominance of the homoduplex $d(GGC)_n$ over hairpin at the DNA concentrations they used. They investigated hairpin folding by electrophoresis in non-denaturing gels. They plotted percentage of hairpin as against duplex DNA for n = 5, 6, 7 and 11 and found that at all salt concentrations tested $d(GGC)_5$ showed a greater tendency to hairpin formation than did $d(GGC)_6$ and attributed this to the number of bases in the loop. The result indicates that if pairing in the stem of the hairpins is the same as that of the duplex, *i.e.* frame 2, then a hairpin 10 structure is preferred over hairpin 9. Another NMR study of these repeats (Zheng *et al.*, 1996) also found in favour of alignment in frame 2. Its deduction is made from observations of $d(CGG)_3$ with the second, third or sixth residue substituted by inosine, with the G residues of the other triplets unsubstituted. With the I2 substitution an imino proton resonance characteristic of an I·C bond was found but with either of the I3 and I6 substitutions the resonance was that of a non-hydrogen-bonded inosine.

The NMR studies agree about the alignment of self-annealing of the G-rich strand but disagree about the mispaired G residues. Zheng *et al.* (1996) found very broad imino-proton resonances for these bases and could not detect amino proton resonances or intra-residue NOEs. They concluded that the residues were unpaired and very mobile, most likely undergoing dynamic exchange among various glycosidic conformational isomers. Mariappan *et al.* (1996b) found an imino-proton peak corresponding to G·G bonds with broad resonances either side which they attributed

to the minor hairpin population (presumably the unpaired G residues in the loop). They also found a strong nuclear Overhauser effect (NOE) connecting the G·C and G·G imino-protons and concluded that the mispaired G residues were strongly base-paired through the imino-protons and stacked with the neighbouring G·C pairs.

It was mentioned in Chapter 4 that Kang *et al.* (1995) found apparent blockages to DNA synthesis through double-stranded d(GCC)·d(GGC) repeats *in vitro* but only if the d(GCC) strand was the template and that Ohshima & Wells (1997) had sequenced the products of this stalled synthesis. As with d(CAG)·d(CTG) and d(GAC)·d(GTC) repeats, these were found to be nascent strands that had folded back onto themselves and they were d(GGC) strands. The template was long (160 repeats with one central interruption) and there were three major products. These were separated from one another in length by only one repeat and contained the sequences d[C(CGG)$_n$] with n = 6, 7 and 8, with the flanking sequences showing that they had formed two odd-membered loops (hairpin 10 in Figure 5.1) and one even-membered one (hairpin 9), *i.e.* the hairpins contained effectively 5, 6 and 7 d(GGC) repeats with d(CC)·d(GG) at the interface with the complementary sequence. Base-modification and cleavage studies showed that the even-membered loop contained four unpaired bases and that the odd-membered loops contained just three, just as in Figure 5.1. The bands in Fig. 3 of Kang *et al.* (1995) that were found by Ohshima & Wells (1997) to correspond to products with 6 and 7 repeats were of approximately equal intensity, suggesting that as with d(GAC)·d(GTC) repeats there may not be nearly as much difference in stability of odd and even loops as found with d(CAG)·d(CTG) repeats.

It was very odd that these looped products were only found with the G-rich strand because, as will be see later in this chapter, all other evidence points to the C-rich strand having a greater tendency to hairpin formation than the G-rich strand.

Investigations interpreted as showing alignment in frame 3

Sinden & Wells (1992) suggested that a single strand of $d(CGG)_n$ might form a hairpin aligned in frame 3 with G·G bonds, quoting papers referring to quadruplex DNA with guanine quartets. Fry & Loeb (1994) examined the possibility of quadruplex formation with short oligonucleotides. They melted them and then incubated for up to 90 hours at 4°C and found that at pH 8 in 200 mM KCl $d(CGG)_4$ and $d(CGG)_5$ would form species that were electrophoretically slowly-moving on non-denaturing gels if the cytosines were methylated and that $d(CGG)_7$ would do so even if not methylated, but such species were not formed by the corresponding C-rich oligonucleotides. They then investigated the dependence of the formation of the slow-moving complexes on the presence of various cations, examined their kinetics and stoichiometry of formation and resistance to methylation by DMS and concluded that the G-rich oligonucleotides formed quadrimolecular quadruplexes. At this time the possibility of quadruplex formation by $d(CGG)_n$ in frames 1 or 2 had not been suggested. The authors assumed that bonding was in frame 3 and none of their experiments could have distinguished the frame. They also apparently assumed that the strands would be parallel (as opposed to antiparallel). Like Lee (1990, who studied other quadruplexes) they found that there was an optimum concentration of Mg^{2+}, 4 mM, for quadruplex formation [by $d(^{5m}CGG)_5$]. The maximum percentage of the total DNA in the slow-moving complex, 12·6 %, was less than achieved with any of the other ions they tried. They plotted percentage of quadruplex formed after a fixed time for different concentrations of K^+, Na^+ and Li^+, and percentage formed at fixed ion concentrations after different time periods. In each case K^+ fostered a much higher percentage than Na^+.

Sen & Gilbert (1990) found that with G-rich oligonucleotides that were capable of forming quadruplexes both from two hairpins and from four strands, K^+ induced the rapid formation of bimolecular quadruplexes that were so stable that they would not unfold to allow quadrimolecular ones to form. To achieve four-stranded

structures it was necessary to use Na^+, which did not stabilize either structure as well, and then K^+ could be substituted after the four-stranded quadruplexes had formed. Thus the finding of better quadruplex formation with K^+ by Fry & Loeb (1994) suggests that either they were measuring bimolecular quadruplex or that the oligonucleotides that they were using had little tendency to form hairpins under their conditions. The latter explanation would agree with the findings of Chen *et al.* (1995), Mariappan *et al.,* (1996b). The surprise was that the greatest percentage of the slow moving complex was reached with Li^+ (about 55% at 400 mM Li^+ in 49 hrs). Fry and Loeb suggested that this might indicate that the guanine tetrads (in these $d(CGG)_n$ quadruplexes) might be packed more tightly than in quadruplexes formed from short guanine tracts dispersed among non-guanine sequences.

Fry and colleagues (Nadel *et al.*, 1995) went on to examine the possibility of unimolecular quadruplex formation by studying fast-moving electrophoretic species. This time, in addition to very short oligonucleotides, they included $d(GCG)_8$, $d(GCG)_{11}$ and other forms of some of these in which one, two or three of the bases were replaced by thymine in each of the places where the unpaired loops would be if the molecule folded into hairpins or quadruplexes. It is clear that they assumed that the alignment would be in frame 3 because the thymines were placed in positions such that only in this alignment would the pairing be unaffected by their presence (whether the structure was a quadruplex or a hairpin). From electrophoretic, kinetic and UV-cross-linking studies they concluded that unimolecular secondary structures were formed. However three pieces of evidence suggested that the structures were hairpins and not quadruplexes. Firstly, their formation was not dependent upon cations. Secondly they concluded that all the guanines were modified by DMS. Thirdly, diethylpyrocarbonate modification (which reveals unpaired purine residues) showed that the $d(GCG)_n$ sequences contained only one loop of unpaired bases.

The authors then deduced the structures of the hairpins from DEPC (diethylpyrocarbonate) and $KMnO_4$ modification results. The results for a substituted form of $d(GCG)_{11}$, $d[(GCG)_2T_3(GCG)_2T_3(GCG)_2T_3(GCG)_2]$ show

minimal cleavage at any of the G residues. If there was folding in frame 1 or 2 there would be some unpaired guanine residues opposite thymines. Thus it appears that this molecule folds in frame 3 with all of the guanine residues in G·G bonds and three loops of unpaired thymines. It is seen however that this is only because the positioning of the thymines is such that it would be energetically disadvantageous to pair in any other way because the unsubstituted $d(GCG)_{11}$ does not pair in this frame. The authors concluded that $d(GCG)_{11}$ paired in frame 1 with a one-base 3' overhang (a hairpin 7 structure) or possibly in frame 2 with a two-base 3' overhang (hairpin 10). This interpretation is marred by the fact that they have interpreted bands on their autoradiograms as cleavage at the wrong ends of the molecules. Reading from the correct end of the molecule, the unpaired loop appears to be 5' GCGG 3'. This corresponds very nicely with the loop of hairpin 9 and indicates frame 2. The loop is an even-membered one and there would be a one-base 5' overhang. The data of Mariappan *et al.* (1996b) suggest that an odd-membered loop may be more stable in frame 2 but with this oligonucleotide this structure would have required a 4-base 5'-overhang or a 2-base 3' one, both of which would probably be energetically less favourable than the one the autoradiogram appears to show.

The reason why Nadel *et al.* (1995) failed to find unimolecular quadruplexes is given by the work of Usdin & Woodford (1995). The latter cloned $d[(CGG)_nC]$ and $d[(GCC)_nG]$ tracts in M13mp18 to make single-stranded templates, then polymerized complementary strands (at pH 9·3) from a primer outside the repeat tract and examined the results by electrophoresis. They found that, with G-rich templates only, there were strong blocks to synthesis of the new strand with all of four polymerases tried. These only occurred if $n \geq 13$. They were at the 3' end of the template so could not be due to formation of triplex DNA between the template and nascent strand. Arrest was independent of template concentration, suggesting an intramolecular structure. For the activity of the polymerase, Mg^{2+} was of course present, as 2·5 mM $MgCl_2$, but the blocks were K^+-dependent. Little if any DNA synthesis arrest was seen in the absence of a univalent cation or when NaCl, NH_4Cl,

RbCl or CsCl were used in place of KCl. With KCl it still occurred even after prolonged incubation of the DNA at 85°C before addition of a heat-stable polymerase. Most of the reactions were carried out in a PCR machine with *Taq* polymerase with 30 cycles through melting annealing and extension, with an extension temperature of 72°C. The other polymerases tried were thermolabile and were incubated at 37°C for 5 min and the same blocks to replication were seen but they were apparently only tried with or without K^+ and not with any of the other ions.

Usdin & Woodford found that arrest of synthesis was eliminated by replacement of the second guanine of each of the last four CGG triplets in a template containing d[$(CGG)_{16}$C] with 7-deazaguanine in which the N7 is not free to take part in hydrogen bonding thus ruling out a hairpin containing only G·C bonds (since these do not involve N7). Since the substitution of only four of the 32 guanines was required to abolish arrest they considered it unlikely that the structure was a hairpin, either with only G·G base pairs (*i.e.* frame 3) or a mixture of C·G and G·G bonds (*i.e.* frame 1 or 2), because the N7 of only 50% or 33% respectively of guanines would have to be involved in these cases.

Electrophoresis and chemical probing was then performed on a 90-mer oligonucleotide containing $d(CGG)_{20}$ and in these experiments no Mg^{2+} was present. In a denaturing gel the molecule ran, as expected, well behind a 69-mer marker. In a non-denaturing polyacrylamide gel containing 0·5 × TBE and no added salt it ran at the same speed as the marker, *i.e.* much faster than before, and under these conditions it would be expected to be a hairpin. In 40 mM LiCl it ran in the same place but in 40 mM KCl it ran faster. This fast mobility was eliminated by methylation of the guanines by DMS. From all this evidence the authors concluded that the structure was some sort of intrastrand quadruplex. They found that in the presence of K^+ ions the oligonucleotide (at the same concentration) was almost completely protected from methylation of the N7 positions by DMS. This is a very impressive result, showing much greater methylation protection than was found by Nadel *et al.* (1995)

with $d(CGG)_{11}$ or Mitas *et al.* (1995b) with $d(CGG)_{15}$. Usdin & Woodford modified their oligomer with BAA followed by formic acid or DMS and then cleaved with pyrrolidine to detect unpaired cytosines and found only the 11th cytosine of $d(CGG)_{20}$ to be unpaired whether potassium was present or not. The authors pointed out that in the absence of potassium this would be consistent with a hairpin 12 structure (Figure 5.1). It could actually fit with hairpins 7, 8, 9 or 10 too (though these would require overhangs). In the presence of K^+ ions, they concluded, the single unpaired C and almost complete protection of all the G residues must indicate a quadruplex in frame 3. This is illustrated in Figure 5.5a (p. 207).

More recently, Chen *et al.* (1998) have also published results of *in vitro* DNA synthesis on single-stranded $d[C(GGC)_n]$ and $d[G(CCG)_n]$ templates in M13. They also produced new strands from these templates in a thermal cycler but they not only used Perkin-Elmer AmpliTaq® polymerase (see Chapter 2), but also the fluorescent-dye-labelled-dideoxy-terminator sequencing reaction mixture containing tris-HCl, pH 9 and an unknown concentration of $MgCl_2$ but no univalent cations. Not surprisingly, they did not find any blockage to replication. With the G-rich strand and n = 21 they found deletion of one and two repeat units respectively in two reactions when they lowered the extension temperature to 45°C, having found no deletion at 60 - 72°C, and with n = 8 there was no deletion even at 45°C. They did not try a polymerase active at 37°C or a constant incubation temperature.

Triad DNA

For completeness it should be mentioned that Kuryavyi & Jovin (1995a,b) proposed a structure for d(CAG) and d(CGG) repeats which they called triad-DNA. In this, the trinucleotide homoduplex forms an antiparallel double helix but alternately two adjacent bases on one strand are paired with one base on the other strand and next to that two bases on the second strand are paired with one on the first strand, the sugar-phosphate backbones having to make unusual turns to achieve this. Molecular mechanics calculations predicted that this structure should be more stable

than a duplex with G·G mismatches. No evidence for the formation of triad-DNA with these repeats has been found though two single (T·A)·A triads have been demonstrated sandwiching two G_4-quartets in a quadruplex of d(TTAGG) pentanucleotides (Kettani *et al.*, 1997).

Studies relevant to the frame of pairing of quadruplexes

The postulation and computer modelling of quadruplexes containing C·G·C·G· quartets by Mitas *et al.* (1995b) has already been mentioned. Two other laboratories discovered that they could actually exist. Leonard *et al.* (1995) crystallized the heptanucleotide d(GCATGCT) and were surprised to find that instead of a Watson-Crick duplex it formed a bimolecular quadruplex composed of two hairpins. These were held together by two G·C·G·C· quartets. In addition, of the bases in the loops, the thymine residues were turned outwards but the adenine residues formed an A·A bond diagonally across the structure. It is notable that this quadruplex has no G_4-quartets. The crystals were formed by vapour diffusion from sitting drops of a starting suspension of about 5 mg/ml of the heptanucleotide in 50 mM Li cacodylate, 50 mM $MgCl_2$.

Kettani *et al.* (1995) set out to investigate the possibility of quadruplex formation by the G-rich strand of d(CGG)·d(CCG) repeats with the oligonucleotide d($GCGGT_3GCGG$). It might have formed hairpins and quadruplexes in any alignment (Figure 5.3, overleaf) but with such a short sequence it is not surprising that the authors found that it formed hairpins with no overhangs, which is actually frame 2. By NMR they found that these hairpins associated to form quadruplexes, using all the G and C bases, with two G·C·G·C· quartets sandwiched between two G_4-quartets. NMR was carried out at pH 6·5 with 0·1 - 0·15 mM NaCl but it took many hours to prepare the sample for examination and there were no observations of how long the quadruplex took to form. While this study showed that a frame 2 quadruplex is possible, the sequence was so short that the energy cost of an overhang might have determined the frame. In the discussion the very interesting statement is

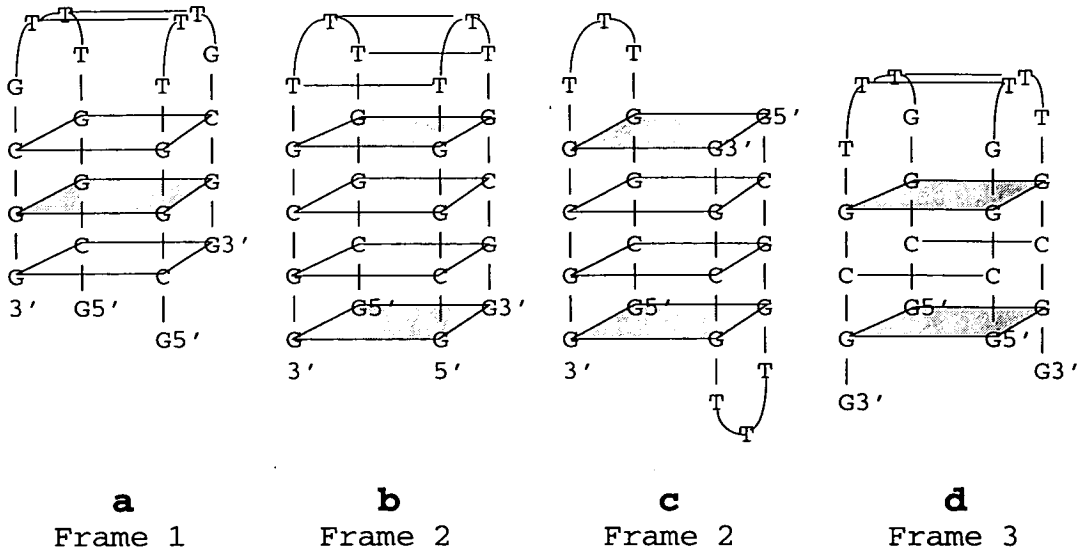


Figure 5.3 Some of the structures that the oligonucleotide of Kettani *et al.* (1995) might have formed. The actual structure was (c) (see text).

---

made that, in contrast to the spectacular proton spectrum of d(GCGGTTTGCGG), the proton spectrum of d(GGCGTTTGGCG) was of poor quality with broad resonances and multiple conformations. If this could be substantiated, it would prove that frame 2 pairing in quadruplexes was preferred over frame 1. Unfortunately, however, there is no mention of this latter molecule in the results section nor any mention of its synthesis in the materials and methods section, nor is there any reference to work on this molecule having been reported elsewhere or having been done but unpublished. To date no NMR study has been published on longer $d(CGG)_n$ quadruplexes.

The C·G·C·G· quartets of these groups are not the same. There are two main types of C·G·C·G· quartet. Both are formed by the association of two C·G base-pairs in opposite directions but in one, which we (Darlow & Leach, 1998a) have called type 1 (Figure 5.4a, overleaf) the major groove sides are facing one-another and in the other, type 2 (Figure 5.4b) the minor groove sides are facing one-another. For the type 1 quartet, different variants have been proposed. In one of these, the two base-pairs are bonded to each other via the N4 of the cytosines and the O6 of the

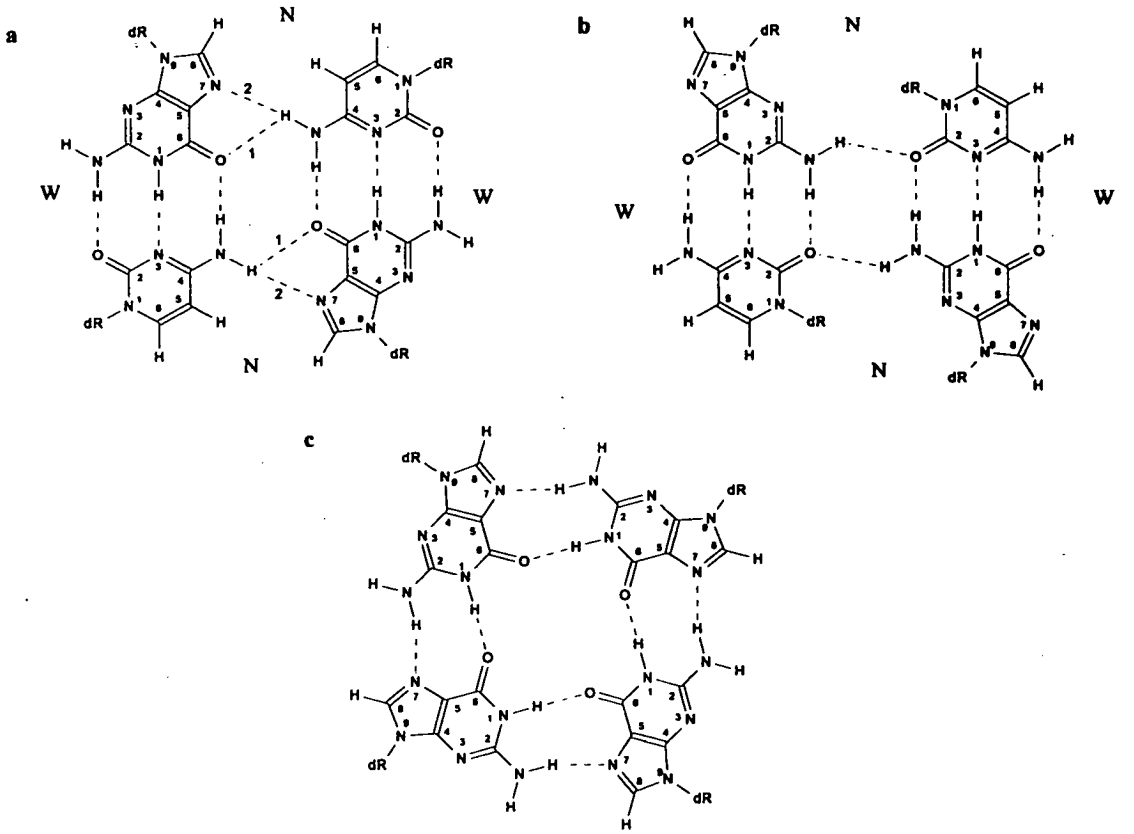


Figure 5.4 a) Type 1 C·G·C·G· quartet. Bonds labelled 1 and 2 are alternative bonding schemes that have been postulated. (b) Type 2 C·G·C·G· quartet. (c) G_4-quartet. W, wide grooves; N, narrow grooves.

---

guanines (bonds labelled 1 in Figure 5.4a). This was the scheme derived by Mitas *et al.* (1995b) from computer modelling and had been previously proposed in the 1960s (Löwdin, 1964; Kubitschek & Henderson, 1966) (for association of two Watson-Crick double helices in models of DNA replication). In this arrangement the N7 positions of the guanines are not involved. Another variation was discovered by O'Brien (1967) by X-ray crystallography of graphite-like crystals of a 1:1 complex of 9-ethylguanine and 1-methylcytosine (no sugar and phosphate involved). In this the N4 of the cytosines were hydrogen-bonded to the N7 of the guanines of the opposite C·G pair (bonds labelled 2 in Figure 5.4a).

Kettani *et al.* (1995) claim that there is a bifurcated hydrogen bond (*i.e.* 1 *and* 2) in the C·G·C·G· quartets in their quadruplex. However, McGavin (1971) drew

this arrangement, half-way between the first two bonding schemes, and commented that the NH---N distance was probably too long for the O'Brien scheme and that the angle between NH and NO was probably too large for the alternative scheme and from his discussion he seemed to consider that the bonding had to be one or the other. Subsequently, Williams *et al.* (1989) have confirmed the finding of O'Brien (1967). Williams *et al.* (1989) used 2'-deoxynucleosides of guanine and cytosine, substituted at both ribosyl hydroxyls with triisopropylsilyl groups and dissolved them in chloroform-*d* a low dielectric solvent which promotes the formation of hydrogen bonds as opposed to stacking. They then examined the result by NMR and, like O'Brien, found C·G·C·G· quartets with bonds 2 only.

The quadruplex of Leonard *et al.* (1995) has type 2 quartets in which the two C·G pairs are bonded to each other via the O2 of the cytosines and the N2 of the guanines. Both type 1 and type 2 quartets have two wide grooves and two narrow grooves but in type 2 the difference between the groove sizes is much larger. Also, the type 1 quartet is roughly planar whereas in the type 2 the two C·G pairs are tilted at ~30° to one-another about an axis going through the C·G bonds. G_4-quartets (Figure 5.4c) are roughly planar and roughly square (though quadruplexes with G_4-quartets do have narrow and wide grooves) so if CGG repeats do form quadruplexes in frame 2 the C·G·C·G· quartets would have to be type 1 in order to stack on the G_4-quartets.

In contrast to the above studies, Chen (1995) found pairing in frame 3. He used absorbance, circular dichroic and gel measurements to monitor the aggregation of $d(CGG)_4$ oligonucleotides beyond a starting mixture of hairpins and linear duplexes. He found that kinetics were extremely slow at pH 8, taking about 10 days to reach equilibrium with 2M KCl at room temperature. A pH of 5·4 and > 0·8 mM KCl were required to observe the onset of aggregation at 20°C within the timespan of 1 day and the formation was of *parallel*-stranded quadruplexes. Chen further concluded that after these formed, the G_4-quartets on either side of the cytosines stacked together leaving the cytosines protruding outwards from the condensed

quadruplex and paired with the cytosines of neighbouring quadruplexes to form larger multiplexes.

Attempts at resolution of the confusion

Several questions arise at this stage: What is the range of conditions under which quadruplexes will form with $d(GGC)_n$ tracts and why did some groups find evidence of quadruplexes and others not? Does the sequence form both frame 3 and frame 2 quadruplexes under different conditions? Does it form quadruplexes under physiological conditions and, if so, what kind are they?

First, why did Mitas *et al.* (1995b) not detect any quadruplex formation with 15 triplets yet Usdin & Woodford (1995) appeared to find it with as few as 13 triplets? The conclusions of Mitas *et al.* hinged upon the methylation conditions but comparison of the methods shows that answer does not lie there. It appears to lie in the annealing conditions. After melting their DNA and immediately before adding DMS Usdin & Woodford incubated at 37 or 55°C for 5 min, during which the quadruplexes evidently formed, while Mitas *et al.* put theirs on ice for 5 min and only hairpins resulted.

It seems likely that Mitas *et al.* (1995b) may indeed have had a monomolecular quadruplex when they estimated the melting point of $d(CGG)_{15}$ secondary structure to be about 75°C. On non-denaturing gels ss-$d(CGG)_{15}$ ran ahead of ss-$d(CTG)_{15}$ suggesting that it formed a more compact structure, possibly a quadruplex. For this work the DNA was pre-annealed at 25°C for 5 min rather than on ice. The gels contained TBE pH 8·5 with no added salt but the DNA samples were diluted in buffer containing 10 mM HEPES at pH 8·5. The buffering range of HEPES is 7·2 - 8·2 (Dawson *et al.*, 1986) so enough alkali, probably NaOH, must have been added to exceed the buffering capacity of the buffer. Thus it might be possible that quadruplexes could have formed when the DNA was mixed with the loading buffer and that they might have been stable enough so that they held the Na^+

ions enclosed in their structure and did not come apart during electrophoresis. Three observations are consistent with this interpretation.

Firstly, though I could not find any reference to a T_m measurement of a unimolecular antiparallel quadruplex of the right size, Sen & Gilbert (1990) found that for tetramolecular parallel-stranded quadruplexes of oligonucleotides of the sequence d(T**GGGGAGCTGGGG**T) the melting point in the presence of K^+ was over 95°C but in the presence of Na^+ only was between 75 and 80°C. This quadruplex had nine G_4-quartets and so might have about the same stability as a quadruplex of $d(CGG)_{15}$ (Figure 5.2). Furthermore, bimolecular antiparallel quadruplexes with 8 G_4-quartets have been reported to have a T_m of about 55°C in Na^+ solution (Hardin *et al.*, 1991) which is all the more reason to think that the structure of Mitas *et al.* was not a hairpin.

Secondly, Mitas *et al.* (1995b) found that when the cytosines of their $d(CGG)_{15}$ oligonucleotide were C5-methylated the melting-point was about 83°C. C5-methylation of cytosines increases the stability of quadruplexes with C·C bonds and G_4-quartets, probably by improving base stacking (Hardin *et al.*, 1993); it increased the stability of the quadruplexes of Fry & Loeb (1994) and appears likely to stabilize quadruplexes with a mixture of C·G·C·G· quartets and G_4-quartets also by improving stacking (Kettani *et al.*, 1995).

Thirdly, Smith *et al.* (1994) produced a supporting result. They performed electrophoresis of $d(CCG)_{15}$ and $d(GGC)_{15}$ and their counterparts with inosine substituted for guanine, $d(CCI)_{15}$ and $d(IIC)_{15}$. [Telomeric sequences with inosine substituted for guanine do not cohere (Henderson *et al.*, 1990; Acevedo *et al.*, 1991).] In a non-denaturing gel, $d(CGG)_{15}$ ran more than twice as far ahead of the I-substituted molecules as did $d(CCG)_{15}$. It seemed possible that this might indicate unimolecular quadruplex formation and Steven Smith (personal communication) kindly supplied the exact experimental details. After melting, the DNA was annealed in 100 mM NaCl, pH 7·4 for 60°C for 10 min and room temperature (22°C) for 10 min. Electrophoresis was performed in TBE pH 8·3 with no added salt but from

their experience they have concluded that when these structures are formed during annealing they are stable during electrophoresis. The absence of smears of label down the gel reassured them about this. They agree with the suspicion of quadruplex formation.

If this is right, it has to be explained why Usdin & Woodford (1995) did not detect blocks to DNA synthesis on a $d(CGG)_{20}$ template in the presence of Na^+. It seems that this was probably because their assay using Na^+ was performed using *Taq* polymerase at 72°C which might be too close to the melting temperature of the quadruplex structure in the presence of Na^+. Chen *et al.* (1998), who found no blockage to synthesis at 45°C had only an unknown concentration of Mg^{2+} (it is not given in the ABI kit information) with neither K^+ or Na^+ added. Fry & Loeb (1994) did find formation of quadruplexes with Mg^{2+} at 4°C but if such quadruplexes with $d(GGC)_{21}$ do not melt below 45°C, they may be too weak to prevent the polymerase progressing through them. Chen *et al.* (1998) do acknowledge that a hairpin or quadruplex may be detected when n is larger or in the presence of KCl (without quoting any source) but they conclude that their work "rules out the formation of G-quartet structure during replication." Of course, this is not so; all cells contain potassium.

We now come to the NMR investigations of $d(GGC)_n$ oligonucleotides. Both studies (Chen *et al.*, 1995; Mariappan *et al.*, 1996b) and (Zheng *et al.*, 1996) looked for evidence of quadruplex formation and did not find it. Both studies used Na^+ rather than K^+, reducing their chances of finding stable quadruplexes. Zheng *et al.* (1996) did most of their work on $d(CGG)_3$ and a UV melting-point study with $d(CGG)_4$ so from the findings of Fry & Loeb (1994) and of Chen (1995) of very slow formation of quadruplex with $d(CGG)_4$ even in the presence of potassium it is not at all surprising that quadruplexes were not found by Zheng *et al.*. Rather, it is surprising that Kettani *et al.* (1995) did find evidence of quadruplex formation by their short oligonucleotide, and with Na^+. Undoubtedly the three thymidine residues play a part. Chen *et al.* (1995) showed that the equilibrium between duplex and

hairpin formation of $d(GGC)_n$ does not go over to mainly hairpin in the presence of 200 mM Na^+ until n >7. Kettani *et al.* found that $dGCGGT_3GCGG$ did form hairpins. The T_3 would be expected readily to form a loop but not to be helpful to duplex formation with this sequence. It is not obvious, however, why two hairpins of $dGCGGT_3GCGG$ should associate to form a quadruplex though two duplexes of $d(CGG)_3$ or $d(CGG)_4$ do not associate readily to form a quadruplex. The sodium concentrations were similar in the studies of Zheng *et al.* and Kettani *et al.* (100 - 150 mM). The DNA concentration used by Kettani et al. (1995), ~10 mM of single strands, was higher than that of Zheng et al. (1996), ~0·6 - 2 mM for 1D NMR studies so perhaps this made the difference. The concentration used by Chen (1995) was very much lower, 40 μM *of nucleotides.*

The other NMR study (Chen *et al.*, 1995; Mariappan *et al.*, 1996b) included $d(GGC)_{3,\ 4,\ 5,\ 6,\ 7\ and\ 11}$, *i.e.* some longer molecules, but most of these investigations were carried out at a much lower salt concentration (only 10 or 20 mM NaCl + 10 mM phosphate buffer) that may not have been high enough for quadruplex formation. (Their DNA concentrations ranged from 0·6 - 3·3 mM.) Identical imino-proton profiles and temperature dependencies were reported for all of the oligonucleotides under all the conditions used but the only test at a higher Na^+ concentration - a salt and temperature-dependent imino-proton profile conducted between 5 mM and 1M NaCl - was again performed on a short molecule, $d(GGC)_5$.

Now we can return to the question of the alignment of folding of the quadruplexes found and what might occur under physiological conditions. Since the publication of our review (Darlow & Leach, 1998a), three more papers have come out (Kettani *et al.*, 1998; Bouaziz *et al.*, 1998; Usdin, 1998) which throw some light on the questions we posed but do not completely settle the matter so I have decided to give the arguments as published and then to discuss the new results.

Fry & Loeb (1994) assumed that the bonding of $d(CGG)_n$ quadruplexes was in frame 3. Usdin & Woodford (1995) deduced it from their results. Chen (1995) showed that a similar base arrangement but with parallel strands can occur, but [at

least with $d(CGG)_4$] only at low pH with very high K^+ concentration and at a very slow rate of formation. In fairness, though, it should be pointed out that this was a tetramolecular reaction with a low DNA concentration. A unimolecular reaction producing a structure with antiparallel strands would be expected to proceed much more rapidly. At the time of their publication, Usdin & Woodford (1995) were unaware of evidence for C·G·C·G· quartets but they have since mentioned (Weitzmann *et al.*, 1996) the possibility that their quadruplexes may contain a mixture of these and G_4-quartets, *i.e.* align in frame 1 or 2. We said that we suspected that bonding may be in frame 2 for the following reasons. Firstly, the stability of the structures of Usdin & Woodford at pH 9·3 and 40 mM KCl makes frame 3 seem unlikely. Secondly, whatever the mode of formation of the tetrahelix, it is likely that the first step would be the pairing of two single-stranded parts of the sequence and the evidence is that duplex pairing is in frame 2 (Mitas *et al.*, 1995b; Chen *et al.*, 1995; Mariappan *et al.*, 1996b; Zheng *et al.*, 1996; Ohshima & Wells, 1997). Therefore, if the quadruplex is in frame 3 then at some time during formation the alignment has to change from frame 2 to 3. Though Usdin & Woodford deduced that in the absence of potassium $d(CGG)_{20}C$ formed a hairpin in frame 3, their chemical modification gels show that the alignment was in frame 2. The bands corresponding to the 5' G of each GpG are more intense than those corresponding to the 3' Gs. This is particularly evident with DMS treatment after BAA treatment and resuspension. At this stage the DNA appears to have been in hairpin form, whatever it had been when initially annealed, and the pattern of guanine bands is just the same as that obtained by Mitas *et al.* (1995b).

Could the results of Usdin & Woodford (1995) support frame 2 bonding for the quadruplex? Alternative structures for $d(CGG)_{20}C$ in frames 3 and 2 are shown in Figure 5.5 (overleaf). In the presence of potassium Usdin & Woodford found almost complete protection of guanines from modification by DMS. If Kettani *et al.* (1995) are right that there are bifurcated hydrogen bonds in the C·G·C·G· quartets then guanines involved in them should be only 50% protected from N7-methylation.

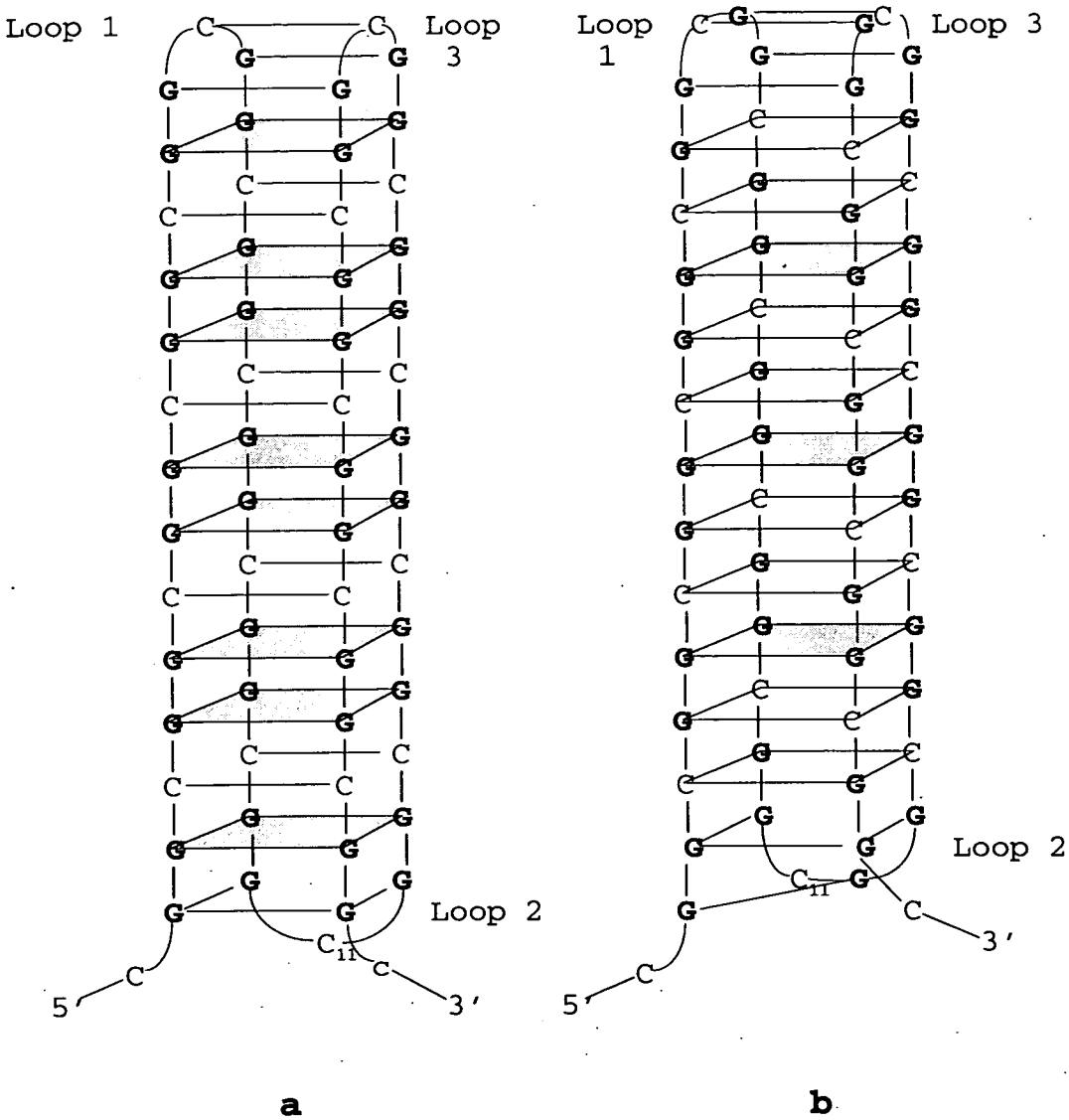


Figure 5.5 Two alternative possible structures for quadruplexes formed by the sequence d[$(CGG)_{20}$C]. (a) One of the frame 3 structures envisaged by Usdin & Woodford (1995); (b) one of the possible structures aligned in frame 2.

However, if as we suspected, the pairing found by O'Brien (1967) and Williams (1989) (bonds 2 in Figure 5.4a) applies, all the guanines in the stem of the quadruplex would be protected, just as they would in frame 3. This still leaves us to explain why bases in the loops are protected, but frame 3 alignment does not appear to explain this either. Usdin & Woodford found only C_{11} (the eleventh cytosine) and no guanines to be unprotected from chemical modification. They deduced that C_{11}

would be in loop 2 (Figure 5.5a) and surmised that the cytosines in loops 1 and 3 might be C·C bonded. This still leaves it unexplained why the loop guanines were not modified.

Usdin and colleagues (Weitzmann *et al.*, 1996) carried out a detailed investigation of the structural requirements of quadruplexes as detected by their polymerase arrest assay and showed that loops may require at least three bases. As we would not expect the four guanines in loops 1 and 3 to be able to form a quartet we should expect that these bases would at best be only 50% protected from modification (two G·G pairs), and we would not expect the guanines in loop 2 to be fully protected either, yet apparently they are all very well protected. In frame 2 (Figure 5.5b) the protection of the cytosines in loops 1 and 3 would be explained by C·G bonding but the guanines involved in these pairs should be completely unprotected and we might expect a further four guanines to be only 50% protected. We also might expect three guanines in loop 2 to be unprotected or only partially protected. However, Usdin and colleagues (Woodford *et al.*, 1994; Howell *et al.*, 1996) showed that the β^A-globin promoter of *Gallus domesticus* apparently forms a quadruplex structure in which all the loop guanines are protected from modification. The authors also suggested that a guanine in loop 2 of this quadruplex might be bonded to guanine of the flanking sequence and this suggestion has been adopted for the frame 2 bonding scheme for d[$(CGG)_{20}$C] (Figure 5.5b). Thus, though we did not know why the loop bases should be so well protected, it did seem that frame 2 bonding could fit just as well with the data as could frame 3 bonding.

An unresolved question for the frame 2 hypothesis was of the ion-binding. Kettani *et al.* (1995) remarked that it was unclear whether univalent cation sites could be generated between adjacent quartets in their quadruplex (bonded in Frame 2) since the internal co-ordination sites now consisted of a mixture of favourable guanine O6 oxygen atoms and unfavourable cytosine N4 amino groups. Another way of putting this might be that internally-binding cations might not be required in this case. We suspected however that a cation might be co-ordinated between a G_4-quartet and a

C·G·C·G· quartet as there would still be six oxygens at these sites. This would mean that two out of every three inter-quartet sites might be occupied by a cation. We did not envisage that a cation would be co-ordinated between two C·G·C·G· quartets. The quadruplex of Leonard *et al.* (1995) had no G_4-quartets. It was not found to have a cation between its C·G·C·G· quartets. Having Type 2 quartets, it had a narrower narrow groove than the quadruplex of Kettani et al. (1995) and was modelled as being stabilized by a Mg^{2+} ion between phosphates. Kettani *et al.* (1995) did not report investigation of the ion-dependence or otherwise of their quadruplex. We only knew that it could exist, at a very high DNA concentration, in the presence of Na^+ and therefore, presumably, larger quadruplexes in Frame 2 could also exist in such conditions, and unimolecular quadruplexes might exist at lower DNA concentrations. Our observations on the work of Mitas *et al.* (1995b) and Smith *et al.* (1994) discussed earlier, as well as the results of Fry & Loeb (1994) argue that $d(CGG)_n$ quadruplexes could exist in Na^+ solution. Therefore it seems possible that they might be bonded in Frame 2 and that their stability might be greatly increased by K^+ binding as we suggested. It is also possible that the structure might be further stabilized by Mg^{2+} in addition to the K^+.

Two of the new papers to which I referred earlier (Kettani *et al.*, 1998; Bouaziz *et al.*, 1998) describe NMR investigations of quadruplexes formed by oligonucleotides with the sequence $d(GGGCT_4GGGC)$ in the presence of Na^+ and K^+ respectively. The oligonucleotide forms a blunt-ended hairpin and in each case a quadruplex is formed from two hairpins with their loops at opposite ends of the structure and running along sides of the structure as opposed to across between diagonally opposite strands. In 100 mM NaCl the quadruplex was found to be just like the one illustrated in Figure 5.3c except that with the new sequence there are two G_4-quartets sandwiched between two C·G·C·G· quartets instead of two C·G·C·G· quartets sandwiched between two G_4-quartets (and four T residues in the loop instead of three). The thymines in the loops are all orientated in different directions. The most 5' is stacked on the preceding cytosine of the same strand. The next

thymine is turned outwards, the next turned towards the interior of the loop, and the last is stacked on the cytosine at the base of the hairpin on the opposite side of the quadruplex. This time the ion co-ordination was studied. The structure co-ordinates three Na^+ ions. One is between the two G_4-quartets. The other two are each between a G_4-quartet and a C·G·C·G· quartet but they are co-ordinated by *seven* oxygen atoms, four from the G_4-quartet, two from the guanines of the C·G·C·G· quartet and one from the inwardly-pointing thymine of the loop.

Thus we still do not know what would be the situation in the stem of a longer quadruplex and, of course, a quadruplex composed of d(GGC) repeats that was aligned in frame 2 would not have any adjacent G_4-quartets but only one G_4-quartet to every two C·G·C·G· quartets.

When the solution suspending the DNA was changed from 100 mM NaCl to 100 mM KCl a conformational change occurred in the quadruplex. The structure is less twisted and the pairs of C·G base-pairs at either end are moved laterally with respect to one-another so that they cannot make linking hydrogen bonds to form C·G·C·G· quartets. (See Figure 5.4a and imagine the base-pair on the left to be moved down the page and the one on the right moved up so that the guanines are level with one-another and then move the cytosines across a bit, the top to the left and the bottom to the right without turning them, so that the hydrogen bonds of the C·G base-pairs are diagonal). The positions of the thymines in the loops are different too. When the authors (Bouaziz *et al.*, 1998) fitted K^+ ions into the structure by computer modelling, they found that five were accommodated. Three were in the same positions that had been occupied by the Na^+ ions except that the ones that had been between a G_4-quartet and a C·G·C·G· quartet partially protruded through between the two C·G base-pairs, and the extra two ions were situated between the loops and an adjacent guanine, co-ordinated by six oxygen atoms, some of them belonging to sugar and phosphate moieties. Brief mention was made that a similar oligonucleotide with one less thymine would form a similar quadruplex in NaCl to the one with four thymines but would not form a quadruplex in KCl solution.

These findings perhaps make a frame 2 $d(GGC)_n$ quadruplex in a K^+ solution less likely than we had thought. It may be that such a structure would form in Na^+ solution but that with K^+ the larger size of the ions would induce a change to frame 3. However, if the alignment is in frame 3 with K^+, it seems very likely that the cytosines would be pushed outwards to collapse the G_4-quartets on either side of them towards each other in order to co-ordinate a K^+ ion between them (look at Figure 5.5a). The finding of co-ordination of ions in loops suggests that this might be a reason for the protection of (at least some of the) bases from modification that was observed by Usdin & Woodford (1995). I hope that Patel and colleagues will go back to their data of Kettani *et al.* (1995) on the $[d(GCGGT_3GCGG)]_2$ quadruplex in NaCl and determine whether Na^+ ions would fit into it and that they will try quadruplex formation with K^+. What is really needed is NMR analysis of a unimolecular $d(GGC)_n$ quadruplex (with no intervening T residues) in a K^+ solution. It would be desirable to have the structure a little longer too, in order to study the stem out of contact with the loops, but perhaps overlapping resonances would make the analysis of a longer structure difficult or impossible with current techniques.

During the preparation of our review I wrote to Dr. Patel to ask him how he and his colleagues arrived at the statement that the hydrogen bonds in the type 1 C·G·C·G· quartets were bifid, where were the data on the d(GGCGTTTGGCG) oligonucleotide that might have formed a quadruplex in frame 1, and what were the time and temperature of resuspension of their lyophilized oligonucleotides during which quadruplex formation occurred. (A copy of the e.mail is included as Appendix 2.) As he did not answer, we had to write the review accordingly, discussing the possible bonding options of the quartet and the conditions of formation of quadruplexes but leaving out mention of the d(GGCGTTTGGCG) oligonucleotide. Our review was published on 9th January, 1998 and the recent papers (Kettani *et al.*, 1998; Bouaziz *et al.*, 1998) were not submitted in revised form till 19th June and though they do not quote us I am sure that the contents of the letter were taken note of because all the points have been addressed in the new papers.

On the first question, Kettani *et al.* (1998) specifically mention the conflicting positions on what the bonding pattern is in the C·G·C·G· quartets and say that whether they started from the position of only that illustrated as bonds 1 in Figure 5.4a or only that of bonds 2, the computer refinements of structure from their data came to the position of bifid hydrogen bonds.

On the second question, Bouaziz *et al.* (1998), like Kettani *et al.* (1995), refer in their discussion to investigations not mentioned in the results sections of either of the new papers. These include the behaviour of the oligonucleotide d(GGGCT$_3$GGGC) in sodium and potassium solutions, but this time, the molecule concerned is mentioned in the methods section of both papers. If then we assume that the results with d(GGCGT$_3$GGCG) mentioned by Kettani *et al.* (1995) without any supporting evidence are true, we have more reason to believe in frame 2 quadruplexes for d(GGC)$_n$ because blunt-ended hairpins of d(GGCGT$_3$GGCG) would be aligned in frame 1 and they were reported not to form stable quadruplexes whereas the frame 2 blunt hairpins of d(GCGGT$_3$GCGG) did do so. It would be interesting to know how the oligonucleotide d(GCGGT$_4$GGCG) or d(CGGCT$_4$CGGC) would behave, because blunt hairpins of these would be aligned in frame 3.

Regarding the third question, Kettani *et al.* (1998) relate that they resuspended six samples of their d(GGGCT$_4$GGGC) oligonucleotide at different concentrations in 100 mM NaCl, 2 mM phosphate, pH 6·6, and incubated them for three weeks at room temperature to ensure equilibration between single-stranded and multimeric states. Measurements then showed a straight-line log-log plot of concentration of quadruplex *v.* concentration of monomeric oligonucleotide, but unfortunately no observations of the times taken to reach equilibrium at the various concentrations are noted.

In the other new paper mentioned earlier, Usdin (1998) concludes that d(CGG)$_n$ forms the same kind of quadruplex as d(TGG)$_n$ and d(AGG)$_n$ and therefore that it is aligned in frame 3. First, she uses her polymerase arrest assay to investigate

synthesis on a $d(AGG)_{20}$ template in the presence of 2·5 mM Mg^{2+} alone or with 50 mM of each of the univalent cations (Li^+, Na^+, K^+, Rb^+ and Cs^+). As in earlier work the assays were performed in a PCR machine at a pH of 9·3. In all cases, very strong blocks are seen around the middle of the tract due to formation of a purine·purine·pyrimidine triplex by folding of the second half of the template back onto the duplex formed from the first half and the newly-synthesized $d(CCT)_n$ strand, but only in the presence of K^+ are there blocks at each trinucleotide at the beginning of the template (*i.e.* 3' end of the repeat), taken to be due to quadruplex formation. She reports that the same results were obtained with $d(UGG)_{20}$ too.

Next Usdin (1998) presents modification and cleavage data for $d(CGG)_{20}$ with DMS, as before (Usdin & Woodford, 1995) but this time taking into account pH, which neither she nor Mitas *et al.* (1995b) previously noted in this investigation. In the presence of 2 mM Mg^{2+} and 50 mM K^+ at pH 9·0, and in the absence of the K^+ at a pH said to be 6·5 in the text and in the figure labelling, but said to be 6·0 in the figure legend and in the methods section, there is seen to be some protection of the second guanine of each GpG but less if any protection of the first guanine of each GpG. The latter are represented by dark bands that are said to be of the same intensity as for guanines outside the repeat. Usdin concludes that these results represent frame 2 hairpins as previously deduced by Mitas *et al.* (1995b). In stark contrast, in the presence of the K^+ and Mg^{2+} at the lower pH (6·0 or 6·5, whichever it really was) both guanines of each pair are very substantially protected, the bands being faint. This clearly indicates a quadruplex. The most extraordinary outcome of these results is that Usdin deduced that the structure in the presence of Mg^{2+} and K^+ at pH 9·0 is a frame 2 hairpin yet she claims that her polymerase arrest assay detects quadruplex formation, and it is carried out at pH 9·3. No comment is made on this inconsistency. I therefore checked the annealing conditions for the DMS modification in the recent paper (Usdin, 1998). The reactions were heated at 95°C for 30 seconds, then 55°C for 30 seconds and then 72°C for 60 minutes. These were the same temperatures used in the 30 cycles of the polymerase assay but in the latter

72°C was also maintained for 30 seconds in each cycle. *Perhaps* a quadruplex is initially formed but in the presence of DMS at 72°C for 60 min it melts and forms a frame 2 hairpin. Certainly the results of Usdin & Woodford (1995) after substituting some guanines with 7-deazaguanine indicated that the polymerase arrest assay does detect quadruplex formation. Usdin & Woodford (1995) also reported that the replication blocks were maintained even after 'prolonged' incubation at 85°C. Why then should the quadruplex melt at 72°C and form a frame 2 hairpin just because DMS was present. The alternative possibility is that the structure modified by DMS was a quadruplex, and that it was aligned in frame 2.

DMS modification and cleavage results for $d(TGG)_{20}$ and $d(AGG)_{20}$ were found to be independent of pH but the evidence for this was not shown. The results at only one, unstated pH were shown. Each molecule was reacted in the absence of any cation, in the presence 2 mM Mg^{2+} alone, and in the presence of Mg^{2+} and 50 mM Na^+ and in the presence of Mg^{2+} and 50 mM K^+. For both oligonucleotides any of the cations afforded substantial protection but the order was $K^+ > Na^+ > Mg^{2+}$ and $d(AGG)_{20}$ was protected better than $d(TGG)_{20}$ and Usdin concluded that the order of stability of the quadruplexes is $d(AGG)_{20} > d(TGG)_{20} > d(CGG)_{20}$.

Two observations suggested that the $d(CGG)_{20}$ quadruplex was aligned in frame 3. Firstly, its stability was pH-sensitive. Usdin (1998) points out that this suggests that the structure is stabilized by protonation. Since the other tetraplexes are said not to be pH-sensitive, this protonation would have to be of the cytosines by a process of elimination, though Usdin herself refers to the pK_a of cytosine relative to guanine. She further mentions that cytosines are protonated at N3 which is involved as a proton acceptor in Watson-Crick C·G bonds and that the result therefore suggests that the structure does not contain these bonds. Instead she suggests that protonated cytosines might stabilize the structure by the formation of C^+·C bonds or *via* increased stacking energy contributions or both.

Secondly, Usdin has tested the relative stabilities of single quadruplex d(GGC) repeat units aligned in frames 2 and 3 with her polymerase arrest assay. She

used an oligonucleotide containing the sequence $d(T_3G_7)_4$, which makes a quadruplex with seven G_4-quartets (and T_3 loops), and two other oligonucleotides, one of which makes a quadruplex in which the central quartet of the seven is replaced by C_4 and the other of which makes one in which two adjacent central quartets are replaced by two C·G·C·G· quartets (stacked as C·G·C·G· upon G·C·G·C· as in frame 2). In the presence of K^+ the polymerase is almost completely blocked immediately by the first quadruplex. The second causes weaker blocks and the third makes the weakest block. Usdin reports that she obtained similar results when she substituted the same guanine residues with thymines and likewise with adenines. Usdin only suggests that this provides indirect support for her frame 3 model, presumably because the alignment is constrained by the outer G_4-quartets, but it is seductive evidence.

There is however evidence of alignment in frame 2. Though the electrophoresis bands from DMS modification and cleavage of $d(CGG)_{20}$ in the presence of K^+ at pH 6 or 6·5 are faint, it can still be seen that in every GpG dinucleotide the cleavage is greater at the first residue than at the second. In contrast, on the gels of $d(TGG)_{20}$ and $d(AGG)_{20}$ the two bands are of equal intensity in every pair, in keeping with their having pairs of G_4-quartets interspersed with T_4 or A_4 layers. Thus it appears that just as at pH 9·0, the alignment of the $d(CGG)_{20}$ quadruplex may be frame 2.

Then there is the evidence of the modification and cleavage of the other bases. Usdin (1998) carried out modification of the thymines with $KMnO_4$ and of the adenines with DEPC and found that both were completely unprotected, even in the presence of K^+ and Mg^{2+}, and concluded that these residues were unbonded, even though we know that T·T bonds are possible and A_4-quartets have been suggested (Lee, 1990). In contrast, Usdin & Woodford (1995) found complete protection of the cytosines of $d(CGG)_{20}$ from modification by BAA. Furthermore, they found this protection to be just as sound when the pH was raised from 6·5 to 9·0, at which pH they say themselves, little if any protonation of cytosines would be expected. The protection of the cytosines could be explained easily by their being involved in C·G

base pairs. Why should the quadruplexes with unpaired A or T residues be more stable than the quadruplex with cytosines if the latter is aligned in the same way and has C^+·C-bonded cytosines? Of course, one still has to explain why the d(CGG) quadruplex should have much less guanine modification at the lower pH. DNA is an acid and melts in alkaline conditions. pH 9 is not very high but I should have liked to see the evidence that only the quadruplex with cytosines was affected by raising the pH.

One other investigation performed by Usdin (1998) was a comparison of the electrophoretic mobilities of oligonucleotides containing $d(NGG)_{20}$ where N = A, T, C, U or an abasic site (symbolised by *). This was done in polyacrylamide with 1 × TBE alone or with 7 M urea or 100 mM KCl. As expected, the greatest mobilities occurred in the presence of KCl. The oligonucleotides with T, U, and * migrated at the same rate as that with C. This confirms that they were all quadruplexes. It does not imply that the alignments of the strands within those structures were the same.

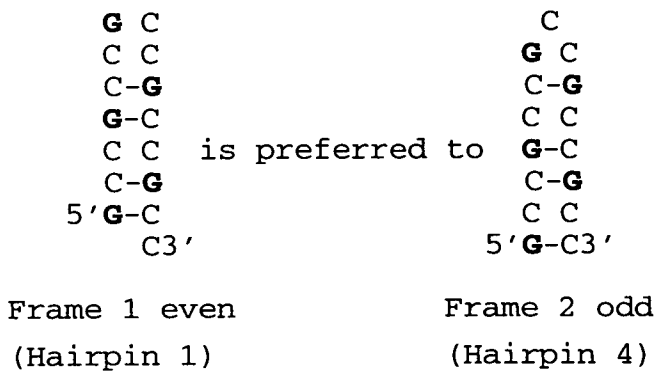
[Interestingly, the oligonucleotide with $d(AGG)_{20}$ migrated more slowly than the rest. Usdin referred to its appearance as a smear and suggested that it might either form a series of different stable conformers with different numbers of repeats or that the tetraplexes were less stable and were continually folding and unfolding during electrophoresis. The latter is rather strange since she had elsewhere concluded that d(AGG) forms the most stable quadruplexes. Actually the appearance on the gel is of two bands, each as broad as but more slowly migrating than those of the other oligonucleotides, and I suggest that they represent different multimers, for instance a bimolecular and a tetramolecular quadruplex. Finally, Usdin (1998) says that her results suggest that expansion disorders involving d(AGG) and d(TGG) repeats remain to be identified. In fact, expansion of both of these repeats in the human genome has already been detected by RED (Lindblad *et al.*, 1994), but both of these repeats are fairly uncommon in genes (Stallings, 1994).]

The C-rich strand: frame 1 or 2?

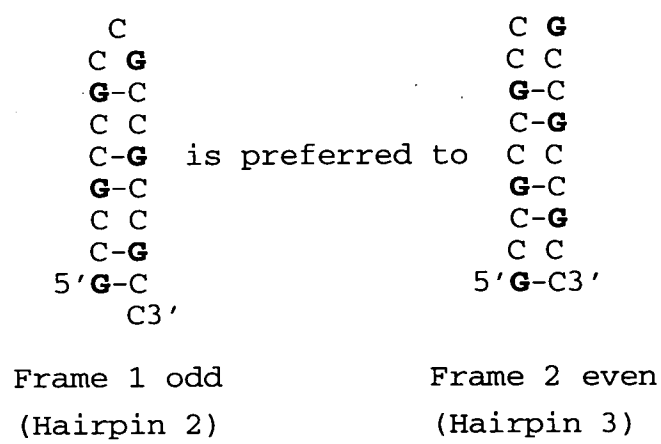
Smith *et al.* (1994) found that in native gels both $d(GGC)_{15}$ and $d(CCG)_{15}$ oligonucleotides migrate faster than non-self-complementary markers, suggesting that they adopt unimolecular secondary structures. They showed that the folded C-rich oligonucleotide was a particularly good substrate for human DNA(cytosine-5)methyltransferase. They seemed to assume hairpin formation in frame 1 as the other frames were not mentioned. In this frame the C of CpG is C·C mispaired, an arrangement in which they suggested that the C would be more easily methylated than if it were in a C·G bond because it would be flipped out of the helix more easily. They found that hairpins of $d(CCG)_{15}$ were methylated about 5 times more rapidly than those in the complementary duplex $d[(GGC)\cdot d(GCC)]_{15}$ and about 8 times more rapidly than those of $d(GGC)_{15}$ in which all the C residues would also be in C·G bonds but bounded on one side by a G·G mispair. They then constrained sequences of $d(CCG)_{11}$ and $d(CGG)_{11}$ to fold in frame 1 by embedding them in flanking sequences which annealed to a complementary oligonucleotide which lacked the sequence to be looped out and again found that the C-rich sequence was a good substrate for the methyltransferase. We noted though that this constrained loop was apparently not as good a substrate as the unconstrained $d(CCG)_{15}$. Further work (Laayoun & Smith, 1995) confirmed and extended information on the methylation of mispaired and unpaired cytosines.

Further electrophoresis of single oligonucleotides (Chen *et al.*, 1995) showed that the C-rich strand forms hairpins much more easily than does the G-rich strand. In 5 mM NaCl hairpin was the predominant form for both strands with 5 - 11 repeats but in 200 mM NaCl $d(GGC)_n$ requires $n > 7$ before hairpin is the dominant form over homoduplex $d[(GGC)\cdot(GGC)]_n$ and there is still an appreciable proportion in the duplex state at $n = 11$ whereas with $d(GCC)_n$ the hairpin is the dominant form even at $n = 5$. The hairpins were then investigated by NMR (Chen *et al.*, 1995; Mariappan *et al.*, 1996b). The results showed that in the C-rich strand the C of the

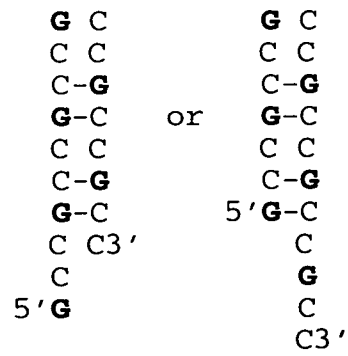
CpG is C·C paired which indicates frame 1. It was found that $d(GCC)_5$ prefers to pair in frame 1 with an even number of unpaired bases in the loop (*i.e.* hairpin 1) and an overhanging 3' C rather than pairing in frame 2 with an odd number of unpaired bases in the loop (*i.e.* hairpin 4) and no overhang, *i.e.*:



$d(GCC)_6$ was also found to pair in frame 1 with a 3' C overhang. In this case the loop has three unpaired bases (hairpin 2 of our scheme), *i.e.* it effectively pairs, as $d[G(CCG)_5CC]$. Since a hairpin with this sequence paired with no overhanging base would be a type 3 hairpin we can infer that hairpin 2 is preferred over hairpin 3, *i.e.*



These investigations with short oligonucleotides did not show whether an odd- or even-membered loop is preferred in the preferred frame 1. In order for $d(GCC)_6$ to form hairpin 1 it would have to have either a 2-base 5' overhang or a 4-base 3' overhang, *i.e.*



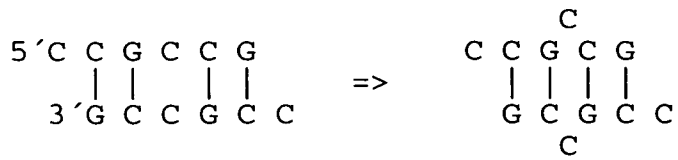
Both Frame 1, even (Hairpin 1).

and these are presumably energetically less favourable, with such a short stem, than having a hairpin 2 structure and only a 1-base 3' overhang. In a genomic setting, bases not involved in the hairpin would not be overhanging but involved in Watson-Crick pairing with a complementary strand so the preferred loop, whichever it is, would have a better chance of forming

Chen *et al.* (1995) also verified the frame of alignments of both strands by the differential cytosine-5-methylation using two different methylases. Both methylate the cytosine of the couplet CpG, but the bacterial methylase *Sss*I requires the cytosine to be Watson-Crick-bonded, *i.e.* it recognizes $\overset{\text{-CpG-}}{\underset{\text{-GpC-}}{\cdot\ \cdot}}$ whereas the human enzyme only requires the guanine to be Watson-Crick-bonded, *i.e.* it recognizes $\overset{\text{-CpG-}}{\underset{\text{-C-}}{\cdot}}$. The human enzyme methylated $d(CCG)_{11}$ at about 6 times the rate of the Watson-Crick duplex, in keeping with the results of Smith *et al.* (1994). The bacterial enzyme, however, methylated it at one third of the rate of the Watson-Crick duplex. If it had been aligned in frame 2 the recognition site for this enzyme would have been present (see Figure 5.1). With $d(GGC)_{11}$ the rate of methylation with the human enzyme was half that of the Watson-Crick duplex, again in keeping with the results of Smith *et al.* (1994) but the rate with the bacterial enzyme was about 1·5 times that of the Watson-Crick duplex. If this strand had been aligned in frame 1 the recognition site for the bacterial enzyme would not have been present.

Another NMR study (Gao *et al.*, 1995) examined $d(CCG)_2$ and several other short C-rich oligonucleotides by NMR to determine the alignment of the

homoduplex. The authors concluded that the alignment of $d(CCG)_2 \cdot d(CCG)_2$ is frame 2 with a 5' overhanging C on each strand and that the single pair of mismatched cytosines do not form a C·C bond but that both Cs protrude outwards (and fold in a 5' direction) *on the same side* into the minor groove with the C·G pairs on either side stacking upon one-another. They named this new kind of duplex DNA 'the *e-motif*'.



However, investigation of duplexes of d(CGCCG), $d(CGC)_2$ and $d(GCC)_2$ showed no evidence of this structure. Spectra of $d(CCG)_{3\text{-}5}$ were interpreted as showing multiconformational equilibria which we suspect might include hairpins. The candidates mentioned were the *e-motif* and parallel duplexes. The spectra of d(CGCCG) and $d(CCG)_{3\text{-}5}$ showed a resonance characteristic of protonated cytosines, thought possibly to indicate $C^+ \cdot C$ bonds. However, the same authors (Zheng *et al.*, 1996) later concluded that in $d(CCG)_{n>2}$ $C^+ \cdot C$ bonds do exist but that the *e-motif* is in equilibrium with a protonated stacked-in form associated with a regular backbone conformation, and that perturbations in backbone conformation associated with anomalous helical structure appear to be dampened by the dynamic motions of the mismatched bases. The study discussed earlier (Mariappan *et al.*, 1996b) did not observe an imino $C^+ \cdot C$ signal. It concluded that the mispaired cytosines stacked within the helix but with a single hydrogen bond (which is possible without protonation) that allows them to flip out of the helix more easily than cytosines in C·G bonds. There thus appeared to be a fair amount of agreement about the mobility of the mispaired cytosines; the disagreement was mainly about which cytosines they are. Zheng *et al.* (1996) suggested that the *e-motif* would favour hypermethylation by 5-methyl transferases because the enzymes require cytosine in an extrahelical position. Unfortunately, the cytosines in the proposed *e-motif*, the cytosines in the sequence GpC, are not the ones that are methylated whereas, as had

been pointed out (Smith *et al.*, 1994; Mariappan *et al.*, 1996b), the cytosines mispaired in frame 1, *i.e.* those of CpG, *are* the ones methylated.

I did not believe in the *e-motif*. The above paragraph expressed disbelief as strongly as we felt able to do in public and after seeing the paper of Yu *et al.* (1997b) discussed below. The disbelief was not only because this pronouncement by Gao *et al.* (1995), Zheng *et al.* (1996) that the C-rich strand aligned in frame 2 was against all the other evidence available up to that time that this strand aligns in frame 1 but because their own data appeared to fit much better with frame 1. This latter point is illustrated in Figure 5.6 (overleaf).

The main reason that Gao *et al.* (1995) gave for believing that $d(CCG)_2$ forms a homoduplex aligned in frame 2 was that they detected an interaction between the guanines in italics in Figure 5.6a. However, as shown in the right-hand column, these bases could also interact in frame 1 alignment when the mispaired cytosines turned outwards. "The *f-motif*" is only a light-hearted name to oppose the *e-motif* and I am not suggesting that the cytosines are necessarily extrahelical but drawing them in this way in this diagram just serves to illustrate the different patterns that are obtained with the two different alignments. It can be seen from Figure 5.6 that $d(CCG)_2$ (a), d(CGCCG) (c) and $d(CGC)_2$ (e) should form very similar structures in frame 2 yet their spectra are quite different and Gao *et al.* (1995) did not explain this. As can be seen from the right-hand column, if the preferred alignment is frame 1, these oligonucleotides would not be expected to be able to form the same structure. I shall resist making further interpretations of this figure but just make the following notes: (i) -xx- represents two triethylene-glycol moieties linking two $d(CCG)_2$ oligonucleotides together; (ii) In the original figure (Figure S4 on Supplemental Page 12) the molecule in (c) was labelled as '$(CGCCG)_2$' but the legend (on Supplemental Page 3) made it plain that the authors actually meant that the structure formed was d(CGCCG)·d(CGCCG) not $d(CGCCG)_2$·$d(CGCCG)_2$.

The molecule on which most of the work of Gao *et al.* (1995) was done was $d(CCG)_2$ and in Table S2 of their supporting information, they say that the C in

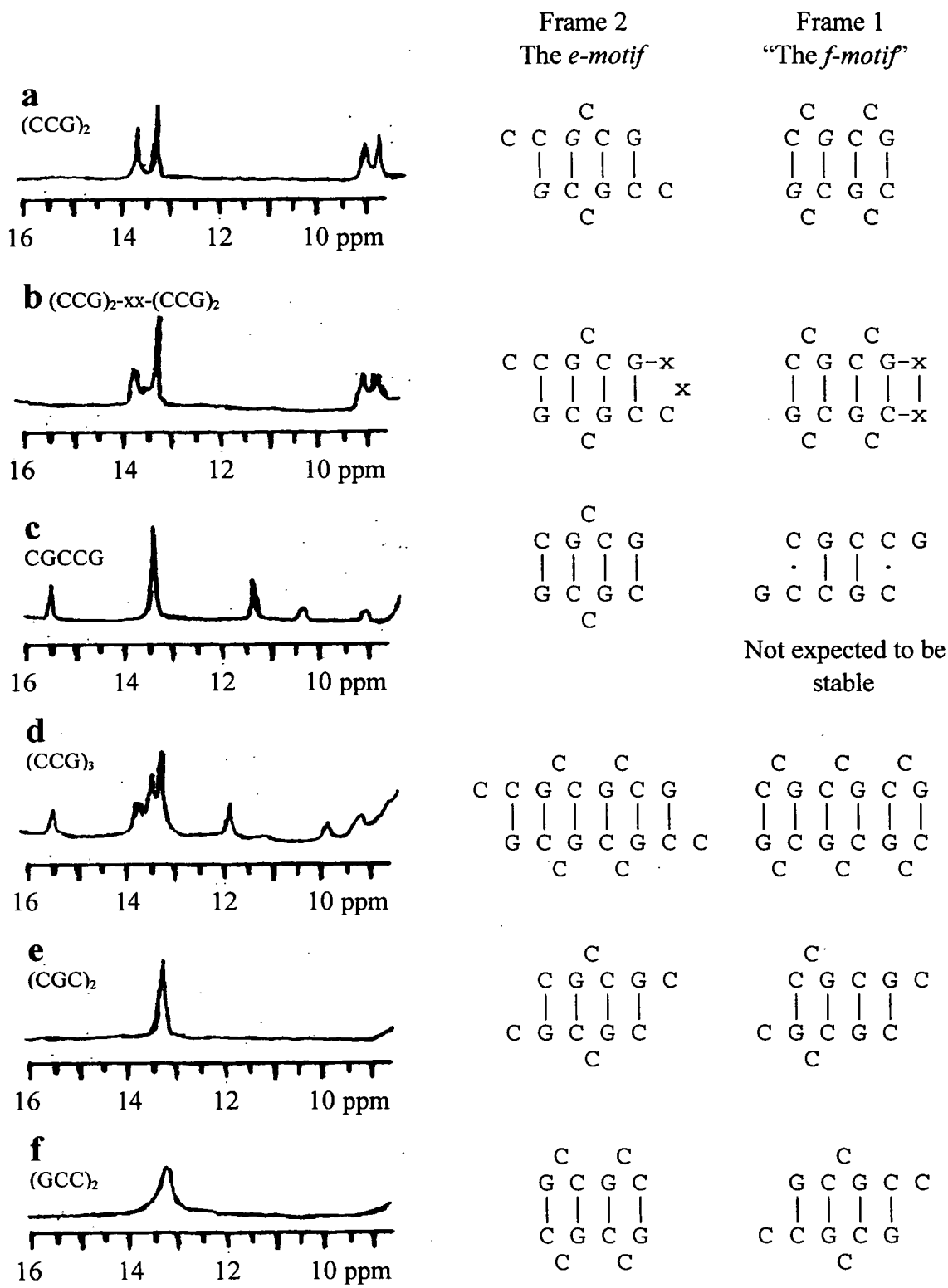


Figure 5.6 1D exchangeable proton resonances of the indicated sequences redrawn from Figure S4 of Gao *et al.* (1995) published in the Supporting Information available free on the Internet under J. Am. Chem. Soc. at http://pubs.acs.org and on microfiche.

position 1 on each chain is stacked with that in position 2 and that the C in position 4 may be hydrogen-bonded (despite the fact that they had proposed that the base was turned out of the helix). Both of these findings fit with frame 1 (with the mispaired bases stacked into the helix) better than with their suggested *e-motif.*

Just as we were about to submit the review (Darlow & Leach, 1998a), another paper (Yu *et al.*, 1997b), came out that brought a considerable surprise and we modified the manuscripts of the review and the paper (Darlow & Leach, 1998b) accordingly. Following work on the G-rich strand (Mitas *et al.*, 1995b), Yu *et al.* (1997b) investigated secondary structure of the C-rich strand by chemical and enzymatic cleavage as well as by physical studies. As before, they considered two alignments, frames 1 and 2, and in these considered quadruplexes and hairpins, and in addition, in both frames, hairpins with all the cytosines turned outwards into the minor groove, both referred to as 'extended e-motif'. First they pointed out that as cytosine has the highest pK_a among all the bases d$(CCG)_n$ might exhibit pH-dependent structural transitions.

As pH was reduced from 8·5 an oligonucleotide containing d$(CCG)_{15}$ (in flanking sequence making a few additional base-pairs) made a sudden increase in electrophoretic mobility (relative to dsDNA) between pH 7·9 and 7·7. The same pattern was seen with an oligonucleotide containing d$(C^mCG)_{15}$ and ones containing d$(CCG)_{18 \text{ and } 20}$ also showed increases in electrophoretic mobility, though over a wider pH range from 8·5 down to the 7·9 - 7·7 region, but oligonucleotides containing d$(GAT)_{15}$ and d$(CTG)_{15}$ showed no such change, the former (not making stable hairpins) moved more slowly and the latter (making stable hairpins) moved more rapidly throughout the range of pH 8·5 - 7·5. Oligonucleotides containing 15 repeats of ATC, CAG, GAC and GTC were also said to show no transition. From this the authors deduced that between pH 8·5 and 7·9 d$(CCG)_{15}$ forms a secondary structure that is more stable than a random coil but not as stable as a hairpin with mismatches and that by pH 7·7 at least some of the cytosines are protonated. No further change occurred on lowering the pH to 6·5. The DNA samples, originally in 200 mM NaCl,

were diluted with 10 mM HEPES by an unspecified amount and were run on polyacrylamide gels made with TBE at various pH values with no added salt at 28°C.

The curve of relative electrophoretic mobility against temperature at ~1 mM Na^+ concentration for $d(CCG)_{15}$ was shifted to the right and steeper with decreasing pH, giving a melting point of 30°C at pH 8·5 but 37°C at pH 7·5, which was taken to indicate modest stabilization by cytosine protonation. The curve of UV absorbance (260 nm) against temperature, in 150 mM NaCl, was similarly shifted and steeper but gave a T_m of 54·C at pH 8·5 and 57·5 at pH 7·5. This was taken to indicate better stacking of the bases when protonated. No comment was made on the large difference in T_m values between methods, but it was obviously due to the sodium concentrations. Circular dichroism spectroscopy (at about 50 mM NaCl) showed two structural transitions, one between pH 8·5 and 6·5 and one between about 6·5 and 4·5. From other spectral features it was concluded that there is better base-stacking with protonation of cytosines at pH 7·5 - 6·5 but not a normal Watson-Crick arrangement and there are no C^+·C bonds at or above pH 6·5 and quadruplexes do not form in this range. Below this pH there was evidence that quadruplexes might be forming with C^+·C bonds between adjacent parallel strands. (In fact, neither the work of Fry & Loeb (1994) or the electrophoresis of Smith *et al.* (1994) or Chen *et al.* (1995) or the DNA synthesis work of Usdin & Woodford (1995) had ever suggested that there would be quadruplexes of d(CCG) repeats at pH values above neutral.)

Yu *et al.* (1997b) then examined the nature of the secondary structure of the oligonucleotide containing $d(CCG)_{15}$ at 37·C at pH 8·5 in 50 mM NaCl by cleavage. This was a pH far above physiological pH and at which the electrophoretic mobility was abnormally slow, whereas by 7·5, the lowest pH at which the authors carried out electrophoresis, the relative mobility was much closer to that of the $d(CTG)_{15}$-containing oligonucleotide. However, it was also clear that stability was increased by higher salt concentration and the circular dichroism had shown that at least there

appeared to be less structural difference between pH 8·5 and 7·5 at 50 mM NaCl than there was at 1 mM NaCl.

Guanines were modified with DMS and unpaired cytosines with hydroxylamine or 2-hydroperoxytetrahydrofuran and cleaved with piperidine. If the alignment was in frame 1 then the cytosines 5' to the guanines would be the mispaired ones but if it was frame 2 then the 3' cytosines would be mispaired. All the guanines cleaved, confirming absence of quadruplex formation. As with their G-rich strand investigations (Mitas *et al.*, 1995b), the DNA was annealed for 5 min on ice but in this case as the other evidence suggested that quadruplexes would be unlikely to form, this was not so important. The very interesting and surprise finding was that the vast majority of the cytosine cleavage was of the residues 3' to the guanines, indicating alignment in frame 2. Cleavage 5' to unpaired bases by P1 nuclease agreed with this. As in the G-rich strand investigations the repeat sequence was set in flanking DNA. The whole sequence of the oligonucleotide was d[GATCC$(CCG)_{15}$G GTACCAAGCT] and it folded to make an even-looped $(GCC)_{14}$ hairpin clamped at its base by $C_3{\cdot}G_3$ with the most 5' C mispaired with a T rather than an odd-looped $(CCG)_{15}$ hairpin which would also have been closed by $C_3{\cdot}G_3$ before the initial C·G base-pair of the repeats. Because the bands corresponding to cleavage of the cytosines in C-C mismatches were as intense or more intense than cleavage of a cytosine mispaired with an adenine outside the repeat region, the authors concluded that there were no hydrogen bonds in C-C mismatches.

At the region which should be the hairpin loop, judging by the alignment indicated by all other cleavages, cleavages of both guanine and cytosine suggest that there was a six-base unpaired loop d(GCCGCC) closed by a 3' C·G 5' from the adjacent repeats. The P1 cleavage results were also compatible with this. Ohshima & Wells (1997) found that the central six bases of an even d(GGC) repeat hairpin, d(GGCGGC), had only four unpaired bases, despite having more bulky guanines. Their reactions were carried out in 50 mM sodium (cacodylate) but at pH 7.

Yu *et al.* (1997b) also found some P1 cleavage in the stem of the hairpin. This was between the bases that would make adjacent C·G base-pairs in frame 2 and from this they deduced that the backbone was distorted as had been proposed by Gao *et al.* (1995). (If the alignment was really in frame 1 this cleavage would be between guanines and mismatched cytosines.) As a control the authors tried P1 cleavage of a hairpin of d(GTC) repeats (whose pairing is analogous to frame 2 of d(CCG) repeats) and found cleavage only in the loop. This was not very convincing because the particular oligonucleotide used had a T residue missing from one of the GTC repeats so that the opposing T had to be looped out, yet there was no cleavage on either side of it with P1 nuclease though Yu *et al.* (1995a) had found the base to be attacked by $KMnO_4$. The authors said that their result indicated that the sugar-phosphate backbone was not distorted but surely it had to be.

The authors then performed the chemical and enzymatic cleavage investigations again at pH 7·5, which they twice referred to as being below neutral. (Even human blood pH is only as high as 7·4 ± 0·4.) They thought that protonation of cytosines might induce a change of alignment to frame 1 with C^+·C base-pairs, even though their circular dichroism results had suggested to them that there were no such base-pairs. The chemical modification results (at 15, 25, 35 and 45°C) indicated that the molecule still aligned in frame 2 and the same results were reported for pH 6·5 at which the cleavages were also tried at 600 mM NaCl. The authors considered that there were still six unpaired bases in the loop at 35°C though I did not find this as convincing on the pH 7·5 as on the pH 8·5 gel. The band intensities did not indicate that there was any point on the stem where cleavage was any greater than anywhere else. However, the P1 cleavage results at pH 6·5 told quite a different story. These indicated that there was very strong cleavage on either side of one particular cytosine in the stem. The same cleavages were recorded at pH 7·5 over ranges of NaCl concentration from 0 - 400 mM and temperature from 37 - 57°C, indicating a very stable configuration. The authors said that there was minor P1 cleavage of all phosphodiesters between this point and the central loop, though no

such cleavage was visible on the photograph in the journal. These cleavages were also reported at pH 6·0, 7·0 and 7·5.

In case the alignment had been forced by pairing in the flanking DNA or by the stability of the loop, Yu *et al.* (1997b) made changes in the flanking sequence and the number of repeats was increased (to diminish the effect of the loop) and changed to an even number (18 and 20 trinucleotides) to change the loop to the one predicted to be more stable in frame 1 (these would have given an even-membered loop in frame 1 but an odd membered loop if aligned in frame 2). Despite all this, the oligonucleotides still appeared to align in frame 2, even looping out a nucleotide at the base of the stem on one side in order to achieve this. P1 nuclease digestion at pH 7·5 also showed increased cleavage on both sides of a single cytosine in the hairpin stem though not nearly as strongly as with the $d(CCG)_{15}$ oligonucleotide.

The conclusion of all these studies was that when $n \geq 15$ d(CCG) repeat hairpins not only pair in frame 2 but adopt the *e-motif* and that the unpaired cytosines that are turned outwards into the minor groove fold back in a 5' direction so far as to stack with another cytosine folded towards it in a 5' direction on the other strand but separated from it by two intervening C·G base-pairs. This model was developed by computer simulation starting from co-ordinates supplied by Gao *et al.* Further simulation predicted that this stacking causes such stress on the helix as to cause an occasional cytosine to be flipped right out. This was in contrast to the conclusion of Zheng *et al.* (1996) from their NMR data that the *e-motif* occurred in duplexes of $d(CCG)_2$ and that, as chain length increased, cytosines were in equilibrium between the *e-motif* and the stacked-in position and backbone distortion was smoothed out, though labile.

The one other possibility raised for the strongly-cleaved cytosine at and below pH 7·5 was that a quadruplex was formed and the base was part of one of the other loops, but this was dismissed for several reasons. Two other possibilities come to mind that they did not mention. One was that the oligonucleotide might not form a single hairpin but two or three, as suggested for d(CAG) repeats by Petruska *et al.*

(1996; 1998), Mariappan *et al.* (1998a). However, though the strongly-cleaved base was about half-way up the stem of the $d(CCG)_{15}$ [*i.e.* $d(GCC)_{14}$)] hairpin, it was nearer to the bottom of the $d(CCG)_{20}$ one. The other possibility was that by looping out a cytosine on one side the rest of the hairpin above that point was able to align in frame 1. However, the chemical cleavage pattern of the cytosines did not support this and in the $d(CCG)_{20}$ hairpin it would be very strange that the cytosine looped out on one side right at the bottom of the repeats would be to allow them to align in frame 2 and then the other one looped out on the other side slightly further up would be to allow realignment in frame 1.

Yu *et al.* (1997b) addressed two questions arising from their results. First, instead of accepting the conclusion of Gao *et al.* (1995), Zheng *et al.* (1996) that short strands of d(CCG) repeats align in frame 2 and concluding that alignment is always in this frame, they accepted the evidence that short $d(GCC)_n$ oligonucleotides (n = 5 - 7) form hairpins in frame 1 (Chen *et al.*, 1995; Mariappan *et al.*, 1996b) and so had to explained how this could be. Their explanation was that with short tracts loop structure and/or end-effects might favour frame 1. Actually, as discussed above, Mariappan *et al.* (1996b) showed that frame 1 was preferred for $d(GCC)_n$ oligonucleotides regardless of whether the loop was odd- or even-membered and even though frame 1 gave 3' overhangs in this frame while frame 2 would have given no overhang. A more plausible explanation might be connected with the fact that both NMR investigations [(Chen *et al.*, 1995; Mariappan *et al.*, 1996b) and (Gao *et al.*, 1995; Zheng *et al.*, 1996)] found that there was some C·C bonding with short oligonucleotides whereas Yu *et al.* (1997b) concluded that there was none with their longer molecules. When the cytosines are in the stacked-in position, frame 1 would be expected to be the more stable because, as pointed out by Yu *et al.* (1997b), the stacking energy of the GpC base-pair steps present in the frame 1 alignment is –14·59 kcal/mol as against –9·69 kcal/mol for the CpG steps present in frame 2. However, with the cytosines turned outwards frame 2 would be expected to be more stable because now the stacking of the base-pairs on either side of the outwardly-turned

cytosines might be more critical and, as pointed out by Yu *et al.* (1997b), this is a pseudo-GpC step in frame 2 but a pseudo-CpG step in frame 1. Furthermore, one might describe the 'extended e-motif' of Yu *et al.* (1997b) as 'locked' by the stacking of cytosines 3 bp apart reaching towards each-other in the minor groove. In short oligonucleotides, perhaps, there might not enough be of these stacked cytosine pairs to stabilize the structure. For instance in a hairpin of $d(GCC)_7$ there could only be two such pairs.

The other question Yu *et al.* (1997b) addressed was of why $d(CCG)_{15}$ hairpins should be a good substrate for 5-methylation of cytosine residues when in frame 2 the wrong cytosines are extrahelical. They proposed two possible solutions. Either there might be a minor population of hairpins in frame 1 or the CpG dinucleotides in the 'extended e-motif' frame 2 hairpin might be an excellent substrate for the human methylase because of the distortion of the backbone, as suggested by results of Laayoun & Smith (1995). (In our review we suggested that the latter might be the explanation to the observation that the data of Smith *et al.* (1994) showed that $d(CCG)_{11}$ constrained to pair in frame 1 was not quite as good a substrate for methylation as the unconstrained $d(CCG)_{15}$ but Chen *et al.* (1998) have come up with another explanation which will be discussed in the final chapter.) We noted that the gels of Yu *et al.* (1997b) did show some cleavage of the cytosines 5' to the guanines consistent with a minor population of hairpins in frame 1.

A question not addressed was of why frame 1 was determined to be the alignment of short oligonucleotides by Chen *et al.* (1995)/Mariappan *et al.* (1996b) but frame 2 by Gao *et al.* (1995)/Zheng *et al.* (1996). The former used oligonucleotides of 5 - 7 trinucleotides that were long enough to form hairpins. They were $d(GCC)_n$ which might have been expected to align in frame 2, yet did not. [$d(CCG)_{5-7}$ and $d(CGC)_{5-7}$ were not examined to see whether they would also adopt frame 1.] The major investigations of Gao *et al.* (1995) were on $d(CCG)_2$ oligonucleotides that were too short to form hairpins so formed duplexes. They might have been expected to align in frame 1, but the data were interpreted as

showing the *e-motif* in frame 2. The investigations of $d(GCC)_2$ and $d(CGC)_2$ were only of the 1D exchangeable proton resonance spectra. They did not support an *e-motif* interpretation but the alignment was not investigated further.

Another question not addressed by Yu *et al.* (1997b) was how a hairpin forming in a single strand of $d(CCG)_n$ would change its alignment from frame 1 to 2 when it reached a threshold length of somewhere between n = 7 and 15. Having concluded that $d(CCG)_n$ aligns in frame 1, we scoured the paper for faults in the authors' conclusion that their oligonucleotides aligned in frame 2 but could not contradict it. We therefore suggested in the paper (Darlow & Leach, 1998b) that outwardly turned cytosines may provide a mechanism by which short hairpins aligned in frame 1, forming in a long tract of repeats, might convert to a frame 2 alignment as more repeats become involved. The first step was for the cytosines mispaired in frame 1 to come out of the helix and for the C·G base-pairs on either side to stack upon one-another. Then we suggested that an extrahelical cytosine might exchange places with a neighbouring cytosine paired to a guanine causing the previously-paired cytosine to be turned out one place further up the stem and that the process might be repeated in a domino-like process up one side of the hairpin and down the other so that change of frame was effected in a chain of little steps. This may be far-fetched but we felt that it was even more far-fetched that a hairpin should suddenly break all its bonds at once to change its alignment just because it had reached a certain length. One of our reviewers said that we did not need to include this in the paper but we felt that the issue had to be confronted. The purpose was to make people think about whether it was really feasible to propose that one alignment applied for short hairpins and another for longer ones under the same conditions.

Now Mariappan *et al.* (1998b) have challenged the conclusions of alignment in frame 2 for any length of d(CCG) repeats. Just as for their NMR investigations of d(CAG) repeats, they have labelled a base, in this case cytosine at N4, with ^{15}N, and, in addition to studying hairpins of $d(GCC)_{5 \text{ and } 11}$ have studied a homoduplex of d(CGCCGCG) which contains the sequence d(GCCGC)·d(GCCGC) which occurs in

homoduplexes of $d(CCG)_n$ aligned in frame 1. They had already concluded from their earlier work (Chen *et al.*, 1995; Mariappan *et al.*, 1996b) that alignment is in this frame for short hairpins and the purpose of studying this duplex was to establish for certain whether the mispaired cytosines (the cytosines of CpG) are hydrogen-bonded, stacked without bonding, or extrahelical. The results confirmed their previous conclusion that the mispaired residues are not only stacked in the helix but are linked by a single hydrogen bond. The signal disappeared above pH 7·66 whereas the G·C bond signal was intact up to pH 8·11 and disappeared above pH 8·5. The temperature at which these observations were made was unfortunately not given. By labelling only certain cytosine residues the authors were able to establish that the same alignment (frame 1) and same cytosine bonding occurs in hairpins of $d(GCC)_5$ and $d(GCC)_{11}$. They found no evidence of protonation of cytosine between pH 6 and 7.

Mariappan *et al.* (1998b) then investigated the pH-induced structural transition reported by Yu *et al.* (1997b). As the pH is reduced the C·G signal in the 1D imino-proton region diminishes and C^+·C and G·G bond signals appear, indicating the formation of a parallel-stranded structure which the authors identified as the i-motif. The i-motif was first reported in two mutually referential papers (Gehring *et al.*, 1993; Leroy *et al.*, 1993) as a structure formed by $d(TC_5)$ and eight other oligonucleotides consisting of dC residues with or without some thymidines but no other bases. The structure is a quadruplex but quite different from the ones so far discussed. Adjacent strands are antiparallel but bonding is not between adjacent strands but between diagonally opposite ones. Thus bonding is between parallel strands with, in the structures first described, C^+·C and T·T bonds and the two parallel stranded duplexes run in opposite directions with their bases intercalated. One might expect that purines would not be able to fit into this structure. Since the initial papers, a number of other reports of the i-motif have come out. The structures may be tetramolecular, bimolecular (formed from two hairpins) or unimolecular. A·A and G·G bonds have been described in i-motif structures (Berger *et al.*, 1995; Gallego

et al., 1997) but only at the ends of the structures and never between both pairs of strands at the same end. The report of Mariappan *et al.* (1998b) appears to be the first that suggests that an i-motif structure can form with G·G bonds between both pairs of strands inside the body of the quadruplex.

At pH 5·6 the frame 1 hairpins and the parallel-stranded structure were in approximately 1:1 proportions and by pH 4·4 the parallel-stranded structure was the dominant form. A similar result was obtained by observing the ^{15}N4-cytosine amino-proton signal. (Even at pH 4·4 the C^+·C imino-proton peaks (which appear in the 14 - 15·5 ppm region) were short and broad; Mariappan *et al.* (1998b) did not comment on the sharp peaks obtained by Gao *et al.* (1995) at pH 6·5 - 6·8 in 100 mM NaCl, 10 mM sodium phosphate, illustrated in Figure 5·6.) Mariappan *et al.* (1998b) repeated investigations of pH-induced transition with $d(GCC)_{6,\ 7,\ 11\ and\ 17}$ and with $d(G^{5m}CC)_{5,\ 6,\ 7,\ 11\ and\ 17}$ and did not find any length-dependent shift in the mid-point of the transition, suggesting that they were all aligned in the same way at pH 7. They state that they chose not to investigate $d(CCG)_2$, studied by Gao *et al.* (1995) since it cannot form a hairpin but maintain that their results unequivocally rule out the possibility of the *e-motif* over a broad range of length and even without the constraints of flanking sequence used by Yu *et al.* (1997b). I should like to believe that they are right. However, they did not work with physiological salt concentrations and used only 10 mM NaCl + 10 mM sodium phosphate. Their results explain the pH-induced changes detected by Yu *et al.* (1997b) but Mariappan *et al.* (1998b) did not try to explain the cleavage patterns obtained after base-modification or with P1 nuclease. They did however point out that if there were a transition from frame 1 to frame 2 alignment at or below n = 15 a sudden change in the methylation rate by the human C5-methyltransferase should occur and that they have found no such change in a series of $d(GCC)_n$ hairpins with n = 5, 6, 7, 10, 11, 15, 18 and 21.

Another point that they could have made is that C5-methylases appear to act by first pushing the target cytosine out of the helix and then methylating it and

returning it to the helix. This pushing-out appears to be achieved by entry of an arm of the enzyme through the minor groove of the helix (Cheng & Blumenthal, 1996). In frame 2 the target cytosine is in a C·G bond and in the 'extended e-motif' access to all C·G bonds from the minor groove is barred by cytosines from mismatches on either side folded across and stacked in front of them in the groove.

Late addition

A few weeks after the publication of the paper of the paper of Mariappan *et al.* (1998b), discussed above, came the paper of Gacy & McMurray (1998) discussed at the end of Chapter 4 and again I have decided to discuss it at the end because it has a bearing upon the structures formed by both strands of d(CGG)·d(CCG) repeats.

In the native gel electrophoresis of oligonucleotides and duplexes of Gacy & McMurray (1998) (in 100 mM NaCl, pH 7) it can be seen that $d(CGG)_{25}$ ran far ahead of $d(CCG)_{25}$, just as could be seen with the respective oligonucleotides of 15 repeats run by Smith *et al.* (1994). However, on the Gacy & McMurray gel there are also $d(CAG)_{25}$ and $d(CTG)_{25}$ and it can be seen that $d(CTG)_{25}$ runs only a little way behind $d(CGG)_{25}$ and $d(CAG)_{25}$ slightly behind $d(CTG)_{25}$ with $d(CCG)_{25}$ a long way behind all of them. Smith *et al.* (1994) ran their oligonucleotides with T_n markers (as well as oligonucleotides with inosine substituted for guanine). The oligonucleotides were of length 45 bp. $d(CGG)_{15}$ ran at the level of about T_{25}; $d(CCG)_{15}$ ran at about T_{32}, but this was still a long way ahead of the T_{45} marker. Gacy & McMurray (1998) ran the duplexes on the same gel and say that they added no dye with their [γ-^{32}P]ATP-end-labelled DNA but ran bromophenol blue markers in separate lanes and from this they calculated the positions at which duplexes, hairpins and unstructured single strands should run on the basis of the work of Maniatis *et al.* (1975). (It is clear that Gacy & McMurray used the data from Figure 9a of Maniatis *et al.* (1975) and this was presumably coupled with the latter's finding that bromophenol blue runs at about the rate of a 65 bp duplex as it seems that the dye was the only marker Gacy & McMurray used.)

On the gel of Gacy & McMurray (1998) $d(CCG)_{25}$ ran only just ahead of the duplexes, which ran, as expected, a little ahead of an unstructured 75 nt single strand. $d(CAG)_{25}$ ran at the position expected for a hairpin with $d(CTG)_{25}$ and $d(CGG)_{25}$ a little ahead. From this, Gacy & McMurray concluded that the latter three oligonucleotides all ran as hairpins and that $d(CCG)_{25}$ did not behave as a hairpin. They noted mention of the *e-motif* by Zheng *et al.* (1996) and of a distorted helix by Yu *et al.* (1997b) and suggested that *either* of these might explain the aberrant mobility, apparently unaware that the latter authors had proposed that their distorted hairpin displayed the *e-motif.*

This contradicts both my conclusion that the G-rich strand might form a quadruplex in the presence of Na^+ and the assertion that the C-rich strand always aligns in frame 1. On the former point, it may be noted that the melting points found by Gacy & McMurray (1998) were in line with those of others. They estimated a T_m of (48·9±0·7)°C for $d(CCG)_{25}$ which was close to the (50·1±0·8)°C for $d(CAG)_{25}$ and (51·4±0·9)°C for $d(CTG)_{25}$, but found a value of (75·1±1·2)°C for $d(CGG)_{25}$, and attributed the differences between the three that they deemed to be hairpins as due solely to differences in hydrogen-bonding and stacking. This seems a little unlikely. The next thought therefore is as to whether a unimolecular quadruplex might only migrate marginally more rapidly than a hairpin with the same number of base-pairs. The quadruplex would only be about half the length of the hairpin but twice as thick. The answer to the question is "Yes". It comes from Figure 7 of Usdin & Woodford (1995). Here they show electrophoresis of an oligonucleotide containing $d(CGG)_{20}$. Unfortunately the authors have mislabelled their oligonucleotides in the figure but one can tell which is which from their mobilities in a denaturing gel. In native gels in the absence of any cation and in the presence of Li^+, in both of which the DNA would only be expected to form a hairpin, the band is level with a marker and in the presence of K^+, in which quadruplex formation was demonstrated, the mobility is a little greater than that of the same marker, but not as far as half-way towards the mobility of another marker 9 nt shorter. Therefore, from the fact that the mobility of

$d(CGG)_{25}$ was found to be a little greater than the mobilities of $d(CAG)_{25}$ and $d(CTG)_{25}$, and the much higher melting point, I am still happy to believe that $d(CGG)_{25}$ does form a quadruplex in 100 mM NaCl.

As to $d(CCG)_{25}$ however, it is a pity that Gacy & McMurray (1998) did not include on their gel a perfect palindrome as a marker hairpin, but since the d(CAG) and d(CTG) strands surely form hairpins, it has to be admitted that the mobility of $d(CCG)_{25}$ is aberrantly slow, even in 100 mM Na^+ at pH 7, and that therefore the C-rich strand may adopt an unusual structure. Furthermore the low mobility would fit with the DNA being rather rigid and curved (Chastain & Sinden, 1998) as the model of Yu *et al.* (1997b) would dictate.

The next question must surely be of whether short homoduplexes of the C-rich strand also show this low mobility but the answer is uncertain. Figure 5 of Zheng *et al.* (1996) shows the mobility of $d(CCG)_4$ in a 20% polyacrylamide gel with 1 × TBE, pH 8·3, with no added salt at 10°C to be slightly less than that of $d(CAG)_4$ and slightly greater than that of $d(CTG)_4$, all of which are well ahead of $d(CGG)_4$. All of the bands are taken by the authors to represent homoduplexes because of their NMR results and if true this would deny a different structure for the homoduplex of $d(CCG)_4$. The authors do acknowledge, however, that the bands are all well ahead of one representing the duplex $d(CGG)_4 \cdot d(CCG)_4$, do not mention at the same time that their NMR measurements were done in ~110 - 210 mM Na^+ at pH 6·3 - 6·8, but do mention that $d(CCG)_4$ undergoes a phase change below 20°C. The absorbance (270 nm)/temperature curve of $d(CCG)_4$ is biphasic. Thus there are two possible reasons why the band of $d(CCG)_4$ might have a similar mobility to those of the other three mentioned even though the duplex structure might be quite different. One is that on this gel they might all be single-stranded. The other is that the other three might be homoduplexes but $d(CCG)_4$ a hairpin. However, the melting-curve suggests that the T_m of the hairpin is about 5°C (and the gel was run at 10°C).

Conclusions

Despite so much work on the structures formed by the single strands of d(CGG)·d(CCG) repeats, questions still remain for both strands. The C-rich strand forms hairpins very readily with as few as five repeat units. I am convinced that all the evidence indicates that alignment of short hairpins and duplexes is in frame 1 with the mismatched cytosines stacked into the helix and linked by a single hydrogen bond not involving protonation. An even-membered loop appears to be more stable in this alignment. There is evidence that hairpins of up to 21 repeat units are aligned in the same manner and yet there is evidence of alignment in frame 2 for hairpins with 15 or more trinucleotides and this has not been satisfactorily explained. The *e-motif* is a computer model and models can be beguiling but incorrect, as witness triad DNA, but the slow mobility of the longer hairpins of the C-rich strand does argue for a structure different from the other $d(CXG)_n$ strands and this might be the structure that Yu *et al.* (1997b) have deduced.

It is agreed that the G-rich strand forms hairpins aligned in frame 2 but these do not form as readily as in the C-rich strand, requiring more repeat units (somewhere above 11) before the homoduplex form becomes undetectable. Mismatched guanines appear to be stacked into the helix and to have at least one hydrogen bond but there is disagreement about this. Odd-membered hairpin loops appear to be a little more stable than even-membered ones and these appear to contain three unpaired bases, GGC closed by a 5' C·G 3' base-pair. It is not agreed whether even-membered loops contain four or six unpaired bases.

The G-rich strand will also undoubtedly form quadruplexes. $d(CGG)_4$ can form parallel-stranded tetramolecular quadruplexes. These have the same base-pairing arrangement as would an antiparallel quadruplex in frame 3 but formation is very slow at room-temperature with > 800 mM K^+ at pH 5·4 and extremely slow at pH 8. However, tracts of 13 or more trinucleotides can form unimolecular (antiparallel) quadruplexes in the presence of Mg^{2+} and only 50 or 100 mM K^+ very

rapidly at 37°C, even at pH 9·3, and these have now been shown to be considerably more stable at a pH in the intranuclear range. Whether these quadruplexes are aligned in frame 3 or frame 2 is still uncertain but chemical modification and cleavage data suggest that it is frame 2. If it is frame 2 it is likely that the structure contains stacked C·G·C·G· and G_4-quartets in the presence of Na^+ but that in the presence of K^+ the C·G·C·G· quartets are splayed apart into two C·G bonds by the larger cation.

Finally, it seems worth quoting the words of Lee (1990), written before the discovery of trinucleotide repeat expansion. "*In vitro*," (he obviously meant "*In vivo*") "Na^+, K^+, Mg^{2+} and Ca^{2+} are all present in the cytoplasm and presumably the nucleus of eukaryotic cells. It is known that the Ca^{2+} concentration, for example, increases dramatically upon fertilization of an oocyte. Therefore the structure which is adopted by the guanine-rich telomers" - and here we might add d(CGG) repeats - "may change during the cell cycle and will be dependent on a subtle balance between the concentrations of Na^+ and K^+ on the one hand and Mg^{2+} and Ca^{2+} on the other." This dependence of quadruplex formation upon cation concentrations could turn out to be a mechanism behind differences in expansion frequency between stages in the life-cycle, tissues and sexes (Ashley & Warren, 1995). As far as I am aware, no report of the effect of calcium upon the stability of secondary structure in d(CGG) repeats has yet been published.

Chapter 6

Laboratory work on secondary structures in d(CGG)·d(CCG) repeats *in vivo*

Introduction

With d(CAG)·d(CTG) repeats there was an obvious alignment for hairpins of the single strands to take, *i.e.* d(CAG)·d(CAG) and d(CTG)·d(CTG), and we started by assuming that all d(CXG) repeats would self-associate in this way, meaning that d(CGG)·d(CCG) single strands would pair as d(CGG)·d(CGG) and d(CCG)·d(CCG), *i.e.* frame 1 for both. Thus it was that after constructing the d(CAG)·d(CTG) and d(GAC)·d(GTC) repeat 'phage discussed in Chapter 4, I immediately constructed a series with $d[(CGG)\cdot(CCG)]_{1\text{-}5}$ in the palindrome centres and plaque size assays of these were completed by mid-January 1995. At that time the only papers to have come out on secondary structure in these repeats *in vitro* were those of Smith *et al.* (1994), who considered the possibility only of frame 1 for hairpins and homoduplexes, and Fry & Loeb (1994) who studied quadruplex formation and naturally assumed that this would be in frame 3 as the only quadruplexes whose structure was known at that time were held together with G_4-quartets, but they did not make any suggestion in that paper about the alignment in hairpins.

A plaque-size assay of the first three 'phage, bearing $d[(CGG)\cdot(CCG)]_{1\text{-}3}$, showed exactly the same pattern as with $d[(CAG)\cdot(CTG)]_{1\text{-}3}$, *i.e.* $d[(CGG)\cdot(CCG)]_2$ made very small plaques, smaller than those with only one trinucleotide, and the 'phage with three trinucleotides made much larger plaques. Several assays of 'phage with $d[(CGG)\cdot(CCG)]_{4 \text{ and } 5}$, or with 3 - 5 repeats, along with the reference 'phage DRL176, gave slightly different answers for the shape of the whole 1 - 5 repeat pattern and I eventually decided that I had to assay all of the 'phage at the same time. The result is shown in Figure 6.1.

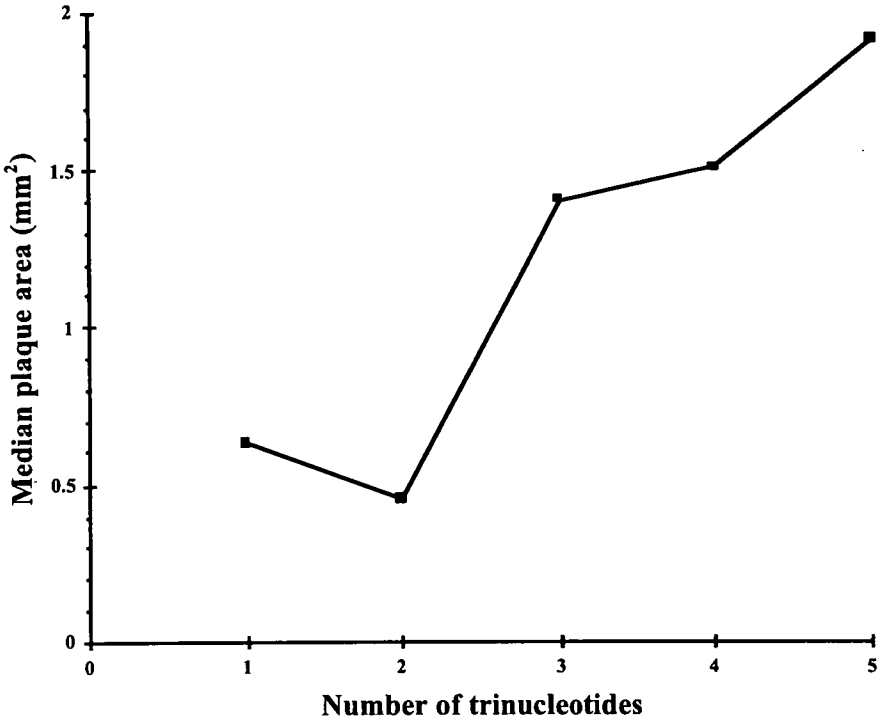


Figure 6.1 The plaque-size results of d(CGG)·d(CCG)-containing 'phage as published in Darlow & Leach (1995).

This pattern was obviously not quite the same as that for the d(CAG)·d(CTG) repeats. I therefore considered the possibility that this might be because one of the strands might prefer a different folding pattern from the one in which it was being constrained to fold. In a long tract of repeats in their normal location in a human chromosome, single strands would be able to fold in whatever alignment was energetically most favourable, but when inserted into the λ 'phage construct (DRL167) hairpins or cruciforms will only form if the palindrome folds in its centre. While d(CAG)·d(CTG) repeats are different from d(GAC)·d(GTC) repeats, d(CGG)·d(CCG) and, d(GGC)·d(GCC) repeats are the same sequence and just because d(GAC)·d(GTC) repeats had never been found to expand I should not take it for granted that one or both of the strands of d(CGG)·d(CCG) might not prefer to fold into a d(GXC)·d(GXC) alignment. There was also the possibility, suggested by Sinden & Wells (1992) that the G-rich strand might prefer to fold so as to make only G·G and C·C, which would be a d(GCG)·d(GCG) alignment. By

inserting the repeats into the palindrome in the way I had done I ensured that the greatest likelihood of a small plaque size being produced would occur if one of the strands had a strong preference to fold in a d(CXG)·d(CXG) alignment (which I named frame 1) because if strands preferred one of the other alignments the fold would never come exactly in the centre of the palindrome whether there were an odd or an even number of repeats inserted. However, by constructing series of 'phage with the repeats in each of the different frames I could check which alignment was most favoured. The only problem might be that, as I suspected from this first set of results, the two strands might prefer to align in different frames from one-another and the double-stranded DNA insert could only be in one particular frame. Nonetheless it was worth trying each of the other frames because if only one of the strands preferred to align in the frame tried it could result in small plaques.

Two more sets of 'phage were therefore constructed, each with 1 - 5 trinucleotides arranged so that folding at the central axis of the palindrome would produce alignment in frame 2 or frame 3. However, first 'phage were constructed and used to investigate the number of unpaired bases in the loops, as described in Chapter 4. Then, because of the difficulty of determining the exact shape of the d(CGG)·d(CCG) repeat plaque size graph, work was done on investigating the sources of error in the plaque size quantification assay and optimizing the assay, as described in Chapter 3. This chapter presents results of plaque assays of all three alignments of d(CGG)·d(CCG) repeats.

Bacteriophage construction and testing

'Phage with inserts containing 1 - 5 of each of d(CGG)·d(CCG) (frame 1), d(GGC)·d(GCC) (frame 2) and d(GCG)·d(CGC) (frame 3) repeats were constructed from λDRL167 exactly as for d(CAG)·d(CTG) and other insert sequences as described in Chapter 4. Because the results of the first plaque assays of the latter two sets of 'phage suggested that some constructs might not have the correct inserts, the oligonucleotides used for construction were checked by Maxam-Gilbert

sequencing and then 'phage constructions were repeated and all selected isolates were checked for insert size as described in Chapter 2. Plaque assays were carried out by the revised protocol described in Chapter 3 and this included a repeat of the frame 1 assay under the same conditions used for the other two alignments.

Results

To give an idea of the appearance of the data from the revised protocol plaque assay, Figure 6.2 (overleaf) shows histograms of the medians of the areas of plaques on every plate in one of the assays, the new assay of the $d(CGG)_n \cdot d(CCG)_n$ 'phage. Each pillar represents the median of the areas of all plaques measured on a single plate which would be disposed in a Gaussian distribution like the one in the histogram in Figure 4.1 (except that that figure shows the measurements from five plates combined). The number of plaques measured is given under each pillar and the total number of plaques measured in the assay was 3,461. Similar numbers were measured in the assays of each of the series of 'phage with the trinucleotides in the other two frames.

Each of the six charts in Figure 6.2 represents a different phage. It can be seen that the differences between the results from different plates are larger for 'phage which produce larger plaques but, were the charts to be printed with different scales so that the heights of the pillars were about the same in each chart, it would be seen that the variation is proportionately quite similar for each 'phage. In each chart the four pillars on the left are the medians of plates from the first stack and the other four from the same positions in the second stack. In each set of four, the left-most represents the plate nearest to the top of the stack for that 'phage and successive pillars represent plates from further down the stack separated by plates dealt out to each of the other five 'phage as described in Chapter 3. With this knowledge it can be seen that much of the plate-to-plate variation derives from the position that the plate

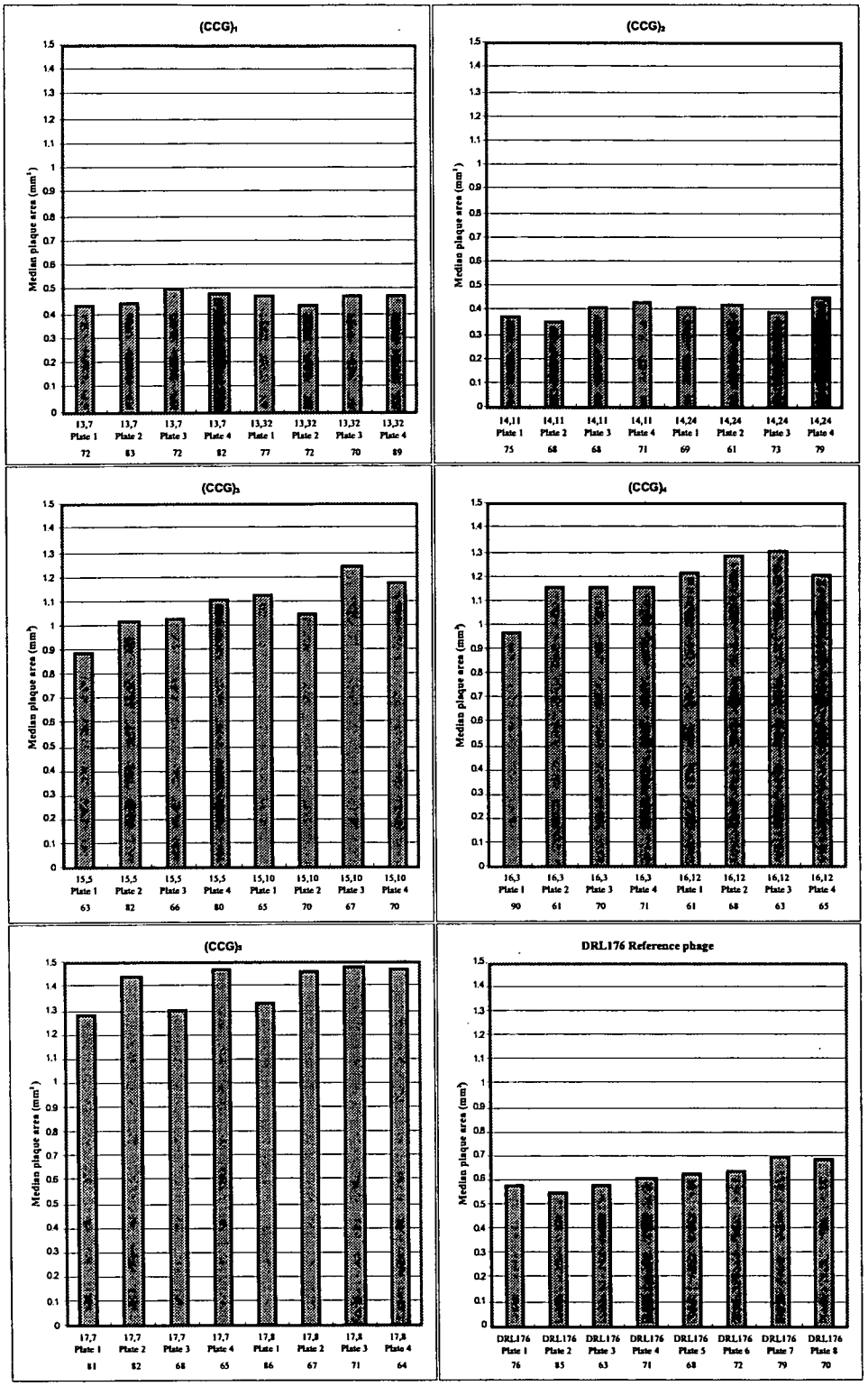


Figure 6.2 Histograms showing medians of areas of plaques measured on all plates used in the final assay of d(CCG)·d(CGG)-bearing 'phage. Columns are labelled with construct number, isolate number, plate number and number of plaques measured.

occupied in its stack during the four days of drying before use. Thus for every 'phage in each stack the average plaque area of the first two plates is smaller than that from the other two plates from further down the stack, though this is less noticeable in some cases than in others. It can also be seen that the plaques on plates from the second stack were slightly larger than those from the first stack, even though all the agar was prepared and autoclaved together and the stacks stood side-by-side during drying. Thus sharing out the plates in the way that I did ensured that false differences between 'phage were not perceived through one 'phage being grown on plates that would tend to produce small plaques and another being grown on ones that would tend to produce larger plaques. I could not totally eliminate sources of error but this seemed to be the best method.

The repeat results for the $d(CGG)_n \cdot d(CCG)_n$ 'phage using the revised protocol gave a very similar appearance to the one shown in Figure 6.1. Figure 6.3 shows the results of the $d(GGC)_n \cdot d(GCC)_n$ 'phage. As mentioned in Chapter 3, the

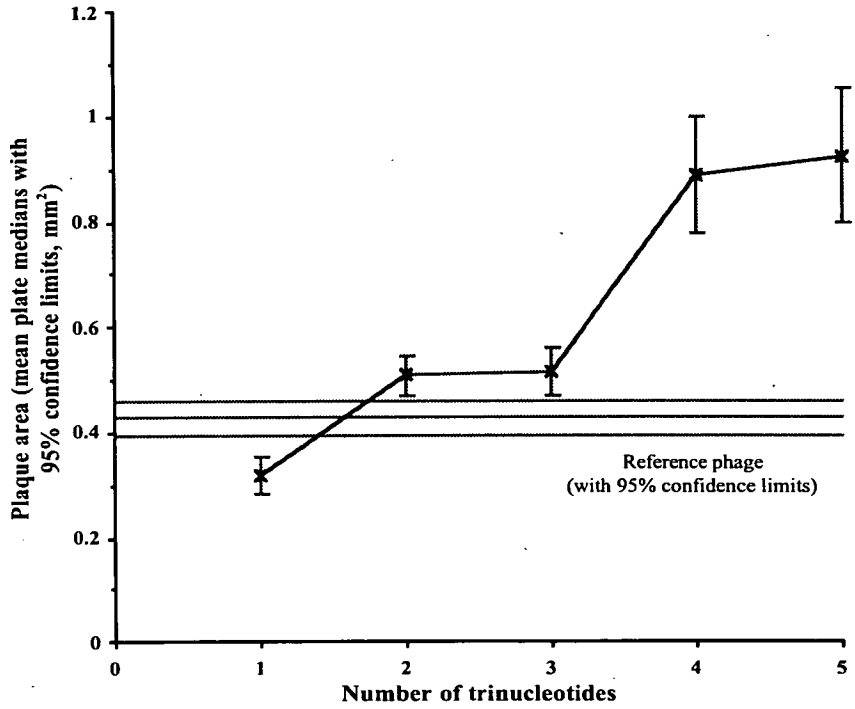


Figure 6.3 Plaque areas of $d(GGC)_n \cdot d(GCC)_n$ 'phage (frame 2) plotted against number of repeat units.

median of the plaques measured on each plate was taken as the best measure of plaque area for the plate and treated as a single result and then the mean was taken of the eight plate medians for each 'phage and the error bars show the 95% confidence limits of those means. Thus the error bars represent the variation of plate medians of plaque area of each 'phage (not the total variation of plaque size which of course was much larger). Figure 6.4 shows the same results plotted with the means of medians from the repeat assay of the $d(CGG)_n \cdot d(CCG)_n$ phage (the data shown in Figure 6.2).

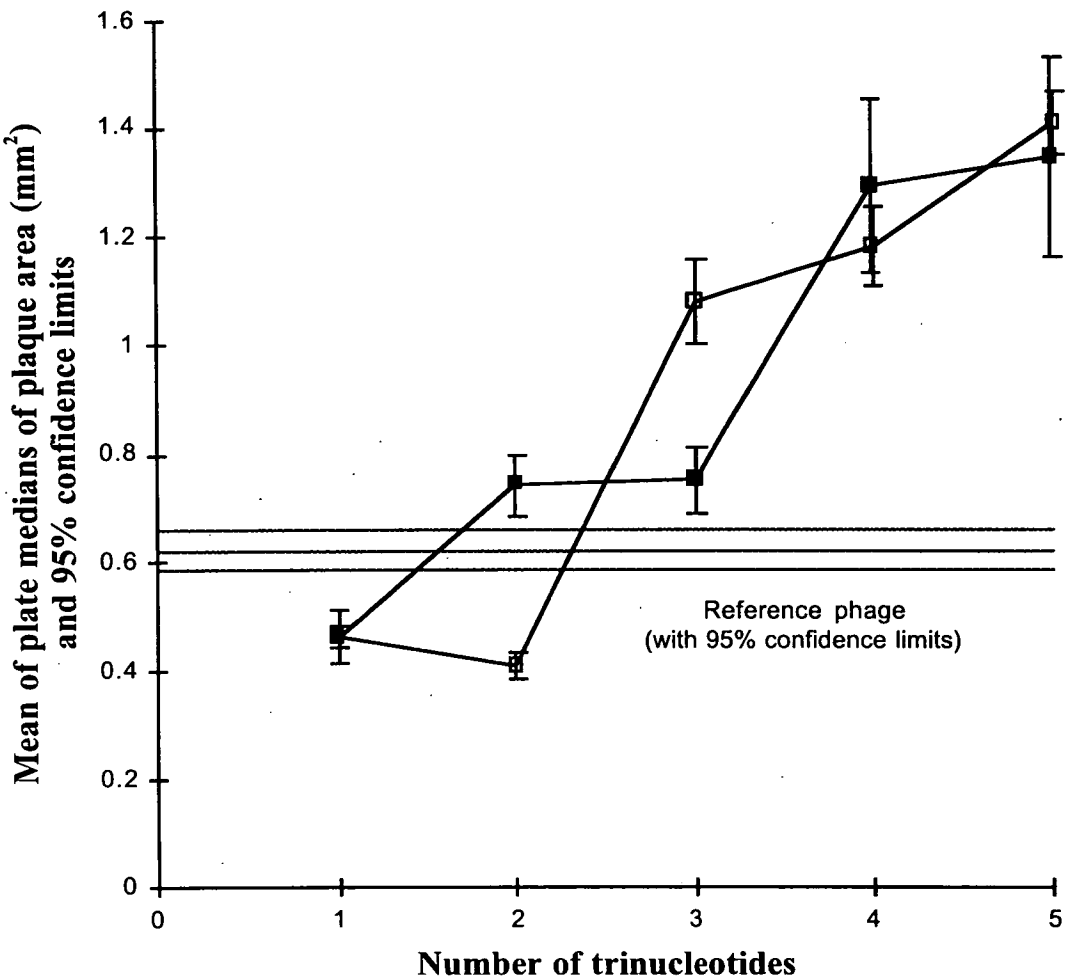


Figure 6.4 Plaque areas of bacteriophage plotted against number of inserted trinucleotides of d(CGG)·d(CCG) (□) and d(GGC)·d(GCC) (■). The overlapping error bars at 4 and 5 repeats have been offset. (From Darlow & Leach, 1998b)

The mean of plate medians of the reference 'phage, DRL176, in the new assay of the d(CGG)·d(CCG) 'phage was 0·62 mm^2 and in the assay of the d(GGC)·d(GCC) 'phage it was 0·43 mm^2, and in Figure 6.4 the d(GGC)·d(GCC) results have been scaled up by the ratio of these values, 0·62/0·43 = 1.46, so that the two plots can be compared directly. It can be seen that the two lines zigzag in opposite directions but come closer together as the number of repeats increases.

Figure 6.5 shows the results for frame 3. Here the line is almost straight.

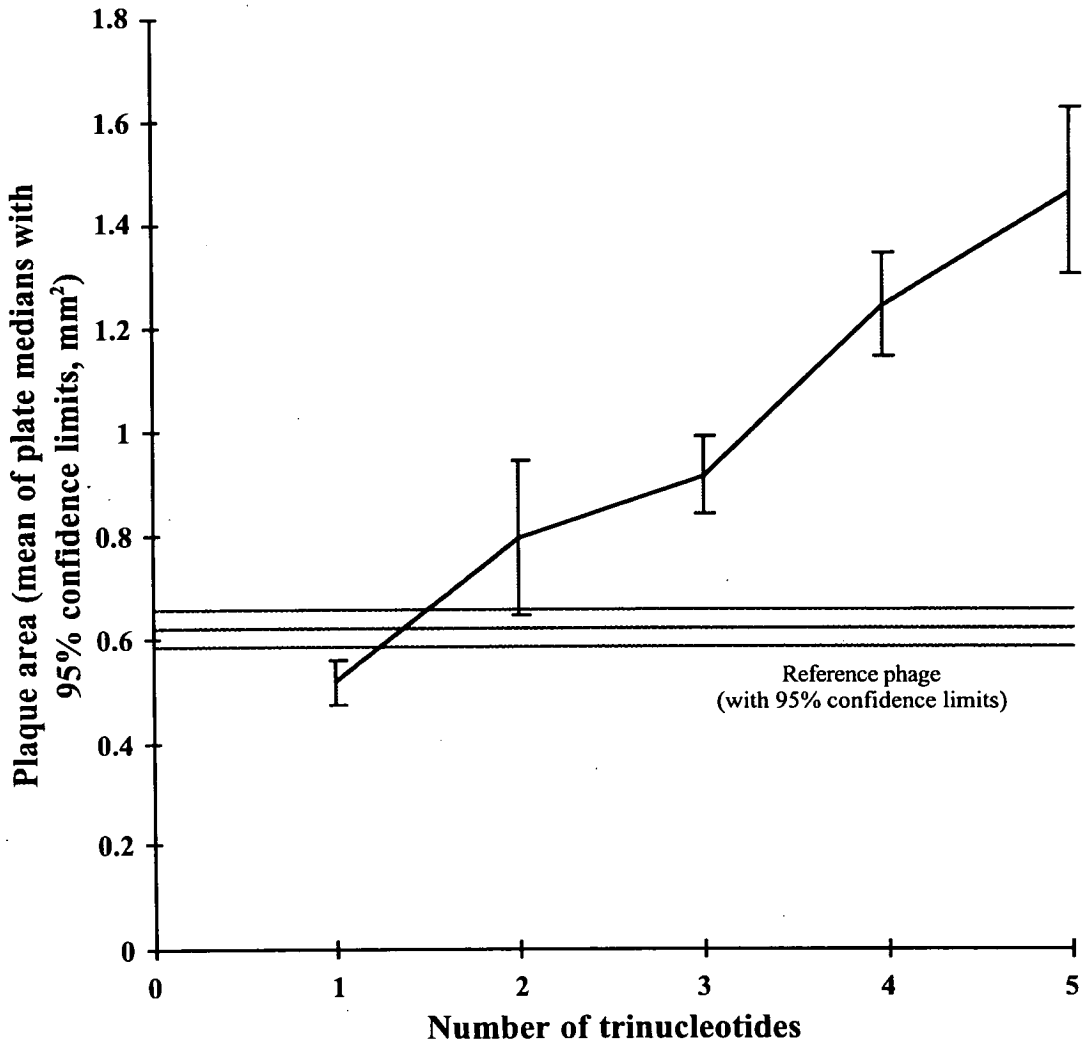


Figure 6.5 Plaque areas of bacteriophage plotted against number of inserted d(GCG)·d(CGC) trinucleotides.

Discussion

Which are the most stable types of hairpin?

Inserts in frames 1 and 2 of odd and even numbers of the trinucleotide repeat investigated here produce alternating patterns of plaque area in opposite directions. As smaller plaques than expected suggest a tendency to form a hairpin with its fold in the centre of the inserted sequence and larger plaques than expected suggest a tendency not to form a hairpin with a central fold, the results suggest that even-membered hairpins are preferred in frame 1 (*i.e.* type 1 or 7 in Figure 5.1) and odd-membered hairpins are preferred in frame 2 (*i.e.* type 4 or 10). In all cases the Watson-Crick base-pair predicted to close the loop of unpaired bases is 5' C·G 3' which is the same arrangement that was found to be favourable with d(CAG)·d(CTG) repeats. (The loops in hairpin types 2, 3, 8 and 9 are closed by 5' G·C 3'.) It is interesting that there is a zigzag pattern in frame 2, $d(GGC)_n \cdot d(GCC)_n$, though there was no such pattern in d(GAC)·d(GTC) repeats, but neither the $d(GGC)_n \cdot d(GCC)_n$, nor the $d(CGG)_n \cdot d(CCG)_n$, plot gives as marked a fluctuation as seen with the d(CAG)·d(CTG) repeats.

From this study alone one cannot tell whether it is the C-rich strand or the G-rich strand that makes the more-stable hairpins in frame 1 and 2 alignments. One might guess that it would be the C-rich strand because the cytosines could stack better into the helix or that it could be the G-rich strand if G·G Hoogsteen bonds form. The frame 1 assay shows that either hairpin 1 is more stable than hairpin 2 or that hairpin 7 is more stable than hairpin 8 (Figure 4.1), the frame 2 assay that either hairpin 4 is more stable than hairpin 3 or hairpin 10 is more stable than hairpin 9.

As discussed in Chapter 5, *in vitro* studies of secondary structures formed by each of the complementary strands have been conflicting. For the C-rich strand there is conflicting evidence from *in vitro* studies of longer repeat stretches of 15 units or more as to whether it aligns in frame 1 or 2 but for short tracts, up to at least 7

trinucleotides, the majority opinion is (and, as I have argued, probably all the results suggest) that it aligns in frame 1. Also, results of Mariappan *et al.,* (1996b) showed and that hairpin 1 is favoured over hairpin 4. Thus the *in vitro* results argue strongly that the result of the frame 1 plaque assay is influenced by the tendency to hairpin formation *in vivo* by the C-rich strand with a preference for hairpin 1 over hairpin 2.

For the G-rich strand the evidence is now overwhelming that hairpins adopt frame 2, as opposed to frame 3, and no evidence suggests that alignment might be in frame 1. Also, an *in vitro* study (Mariappan *et al.*, 1996b) showed that hairpin 10 is favoured over hairpin 9. This would fit with the result of the frame 2 assay being influenced by hairpin formation by the G-rich strand. However, Chen *et al.* (1995) showed by electrophoresis of suspensions of single oligonucleotides that when annealed in 200 mM NaCl $d(GGC)_n$ requires $n > 7$ before hairpin is the dominant form over homoduplex $d(GGC)_n \cdot d(GGC)_n$ and there is still an appreciable proportion in the duplex state at $n = 11$ whereas with $d(GCC)_n$ the hairpin is overwhelmingly the dominant form even at $n = 5$. Thus it seems that neither strand immediately forms hairpins in frame 2 with small numbers of trinucleotides.

The experiment of Chen *et al.* (1995) mentioned above showed that for a short $d(GGC)_n$ oligonucleotide, formation of a homoduplex $d(GGC)_n \cdot d(GGC)_n$ is energetically more favourable than formation of a hairpin with the same mismatching but less than half as many bonds. In the *in vivo* system that I have used the situation is different. The question is: after melting of a short complementary $d(GGC)_n \cdot d(GCC)_n$ sequence, which strand is more likely to form a frame 2 hairpin that will be stable long enough for the flanking perfect inverted repeat sequences to start coming together to form a perfectly-matched stem that will support the fledgling imperfect hairpin? In our paper (Darlow & Leach, 1998b) we suggested that probably again it is the C-rich strand that has the greater tendency to hairpin formation, hairpin 4 being favoured over hairpin 3. However, as discussed in Chapter 5, Ohshima & Wells (1997) found that products of stalled *in vitro* DNA synthesis showed that nascent strands of d(GGC) repeats had folded back onto themselves to

form hairpins with very few repeats that had been stable long enough to be caught by the polymerase and extended to form a complementary stem just like the complementary stem in the extruded palindrome.

Thus it is possible that the reason that the plaque assay results for the repeats in frame 1, d(CGG)·d(CCG) do not give exactly the same pattern as for d(CAG)·d(CTG) repeats is that even with small numbers of repeats the two strands prefer to form hairpins in different frames. Otherwise, if we say that with such small numbers of repeats only the C-rich strand is effective in forming hairpin loops, it might seem to be that above three repeat units this strand must be almost as likely to align in frame 2 as in frame 1. I say this because several plaque assays were done with the d(CAG)·d(CTG) repeats and with the d(CGG)·d(CCG) repeats and there is no doubt that the plaque size for $d[(CAG)\cdot d(CTG)]_4$ is much smaller than that for $d[(CAG)\cdot d(CTG)]_3$ but that the plaque size for $d[(CGG)\cdot d(CCG)]_4$ is larger than that for $d[(CGG)\cdot d(CCG)]_3$ and something has to account for this. An alternative explanation is that the small size of the $d[(CAG)\cdot d(CTG)]_4$ plaques results from the strong tendency of even-loop formation by the CTG strand being augmented by a weaker tendency of the CAG strand also to form an even loop whereas the larger size of the $d[(CGG)\cdot d(CCG)]_4$ plaques results from the strong tendency of the CCG strand to form even loop not being augmented by any particular folding tendency in the other strand. One would then have to explain the frame 2 result by saying that if aligned in frame 2 the $d(CCG)_n$ strand only has a little greater tendency to form an odd loop than an even loop rather than explaining the $d[(GGC)\cdot d(GCC)]_3$ result by saying that the plaque size would be smaller than for $d[(GGC)\cdot d(GCC)]_2$ because odd loops are preferred in the G-rich strand in this frame but the C-rich strand frustrates this by tending to form an even loop off centre in frame 1. There is no *in vitro* data that can help with this.

The fact that the results for frame 3, illustrated in Figure 6.5, plot to almost a straight line even with small numbers of repeats suggests that there is not much tendency to hairpin formation in this alignment. It does not, however, rule out the

possibility that long tracts of the G-rich strand could form stable secondary structures in this alignment, in particular quadruplexes, though, as discussed in Chapter 5, evidence for such structures might be explained by folding in frame 2.

What happens in longer tracts?

With 'phage with all three frames of d(GGC)·d(GCC) repeats, as for $d(CAG)_n \cdot d(CTG)_n$ and $d(GAC)_n \cdot d(GTC)_n$ 'phage, increasing length of non-palindromic DNA between the two inverted repeats that constitute the long palindrome lead to an increase in plaque size. As discussed in Chapter 4, previous work suggested that it appears to be the stability of the protocruciform that is important in cruciform extrusion *in vivo* (Davison & Leach, 1994a) and the plaque assay is most useful to identify the folding position(s) that lead to the formation of the most stable quasi-hairpin(s) formed from a small number of trinucleotides. A small quasi-hairpin could be a nucleating structure that could extend to form more complicated secondary structures in longer tracts of the repeats just as a protocruciform can extend to form a much larger cruciform in a palindromic sequence. The general tendency for plaque size to increase with increasing length of trinucleotide repeat tract inserted suggests that the assay may not be suitable for detection of larger structures. This may not only be because increasing numbers of repeats bring increasing numbers of off-centre copies of the sequence that can form the preferred hairpin-loop but because the stability of the trinucleotide repeat quasi-hairpins may be lower than fully base-paired hairpins and the duplex-hairpin equilibrium may be less favourable to hairpins for trinucleotide repeats than for palindromes. Thus the observations therefore, do not argue against the formation of large secondary structures in long arrays of trinucleotide repeats.

Zheng *et al.* (1996) suggested that the high folding propensity and dynamic properties which they found in d(CXG) repeats *in vitro* should facilitate formation of local structures, not necessarily hairpins, in competition with a linear duplex in genes

containing these repeats. Wells (1996) drew diagrams of looped-out structures in different places on the two strands of a trinucleotide repeat tract which he saw as the only reasonable explanation for the deletion behaviour of such tracts in mismatch-repair deficient bacteria in his laboratory. He called these arrangements 'slipped structures'. As discussed in Chapter 4, Pearson & Sinden (1996; 1998a) have shown that multiple alternative structures ('S-DNA') do indeed form in complementary duplex DNA *in vitro* when trinucleotide repeat tracts of disease-causing lengths are melted and reannealed and have drawn similar diagrams of possible structures. It has yet to be settled conclusively whether long tracts of the C-rich strand of d(CGG)·d(CCG) repeats align in frame 1 or frame 2. It is even possible that when unconstrained by complementary flanking sequences they may not form long hairpins at all but prefer to form multiple smaller hairpins, though this would still, somehow, have to confer a lower polyacrylamide-gel-migration-rate than that of any such arrangement formed by other d(CXG) repeats. However, the preference of shorter tracts to align in different frames, meaning that they would not form hairpins exactly opposite one-another, might help to generate S-DNA structures.

Chapter 7

Construction and testing of a 'phage which allows predetermination of insert orientation, sequencing across the palindrome centre in its context, and introduction of degenerate inserts

Introduction

The main problems with using λDRL167 to test sequences were that it was not possible to predetermine the orientation of an insert, it was not possible after insertion to determine in which orientation the sequence had inserted, and it was not simple to check that the insert had the intended sequence.

By convention, the λ 'phage genome has a left end and a right end and a top strand and a bottom strand and these terms will be used in discussion of constructs made from it. The insertion site in DRL167 is a palindromic restriction site (of *Sac*I) in the centre of a long palindromic sequence (in which 5'→3' the two strands are the same) but the inserts had two different strands. It was of course not possible to predetermine whether an insert containing, for instance, d(CAG)·d(CTG) repeats would ligate with the CAG or the CTG repeats on the top strand. It would go into the symmetrical site sometimes one way and sometimes the other. However, my predecessor, Angus Davison found that secondary structure in the 462 bp palindrome made sequencing across it impossible. The palindrome has, not far from its centre, a *Taq*I site - on both sides of course (see Appendix 1) - and he used this to excise the short central region of the palindrome containing the insert and subcloned this fragment and was then able to sequence it. However this could only confirm that the correct sequence had been inserted; the orientation was lost.

In both Angus' work and my own reported so far, whenever plaque assays were carried out on two isolates of a 'phage the results were similar. This could have

been because the orientation of the insert with respect to the origin of replication had little or no effect on the result of this assay, or that flipping of the centre due to recombination between the arms of the palindrome was so frequent that every 'phage preparation - however many times plaque-purified - contained a mixture of 'phage with both orientations, or it could have been that with the insert in one orientation the 'phage was inviable and all isolates had therefore the same orientation. We expected that the first was true because of a belief that the size of the plaques depended upon the likelihood of palindrome extrusion, which should be independent of the orientation of its centre, but DRL167 did not give the opportunity to establish this, nor to investigate the frequency of flipping of the orientation of the palindrome centre.

Another purpose for which the method used so far was unsuitable was the construction of a library of 'phage containing in the palindrome centre all possible sequences of a given length for screening for other sequences that strongly promoted cruciform extrusion. The idea of wanting to be able to do this came from the realization that construction of series of 'phage to test every one of the possible trinucleotide repeats (let alone di- and tetra-nucleotides too) in each frame for possible hairpin promotion would be rather time-consuming and expensive. Construction of 'phage with random centres to the palindrome would allow screening for isolates that produced small plaques on PSQ agar. The trouble was that the inserts made for DRL167 were produced by annealing two complementary oligonucleotides. What was needed to test random sequences was a 'phage that would allow insertion of a single degenerate oligonucleotide and the filling-in of the opposite strand by polymerase. This would require a 'phage with a palindrome which contained at its centre two restriction sites with opposite orientations. Also, after picking small plaques and replating to plaque purify, and to weed out 'phage that had just produced a small plaque the first time because of late adsorption of 'phage onto bacteria, the palindrome centres would need to be sequenced.

What was needed therefore was a 'phage containing a long palindrome which had near its centre two restriction sites giving opposite overhangs. At these positions it would have to be asymmetric, otherwise either restriction enzyme could cleave on either side, but it would have to be compatible with restoration of complete symmetry when a new insert was ligated between the sites. Thus one could create 'phage which would have perfect inverted repeats surrounding the central test sequence of the new insert yet know the orientation of the insert. In order to sequence the centre it would be necessary to have a single base-pair of asymmetry between the two sides so that one could re-cleave the palindrome on only one side of the centre and sequence the arm bearing the insert, but this asymmetry could be in the new insert itself. It would also be necessary to determine how much effect such asymmetry had upon the plaque size by testing the same sequences within the context of a perfect and a slightly asymmetric palindrome.

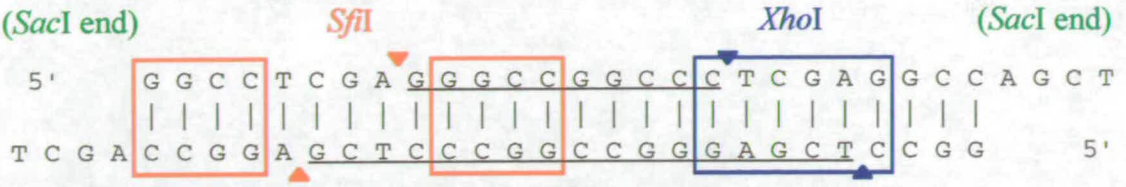
This chapter describes the design and construction of such a 'phage, the identification of isolates with their centres in opposite orientations and a test of whether flipping of the centre is very frequent. It then goes on to describe investigation of the effects of orientation and of asymmetry on a pattern of plaque sizes already observed in DRL167. Finally, results are presented on a series of 'phage bearing d(GAA)·d(TTC) repeats.

The design

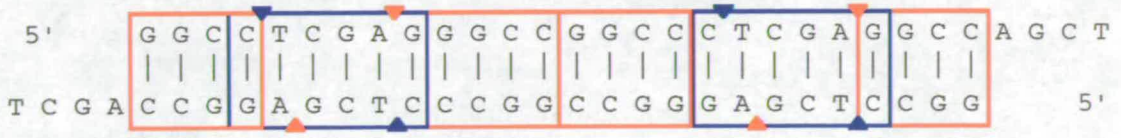
i) The new parent 'phage

The new 'phage was constructed from DRL167 in the same way as all the others made so far, by inserting DNA into the *Sac*I site at the centre of the palindrome and destroying the *Sac*I site in the process. The new restriction sites introduced both had to be ones not found in the λ genome. One of the ones chosen was that of *Sfi*I because it is d(GGCCNNNN/NGGCC) (both strands, the oblique line representing the cleavage position) which gives the possibility of tailoring the

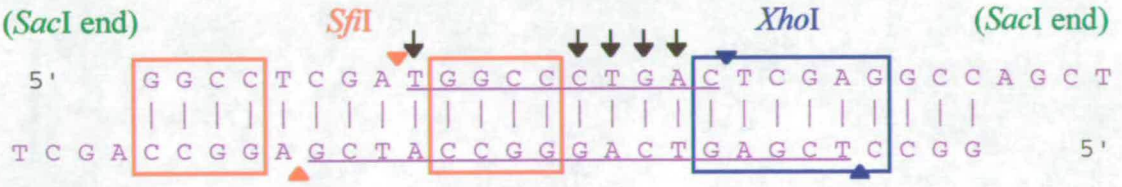
undefined region to suit the other restriction site. The other was that of *Xho*I which is d(C/TCGAG). The minimum palindromically symmetrical insert that could incorporate these sites and ligate into and destroy the *Sac*I site [d(GAGCT/C)] would be:



but, being symmetrical, this has both sites on both sides:



One could then arrange for the removal of one site on each side by making asymmetric base changes confined to the region which would be excised so that a new insert could restore symmetry. The base-pairs marked have been changed from the version above. The segment to be excised is underlined:



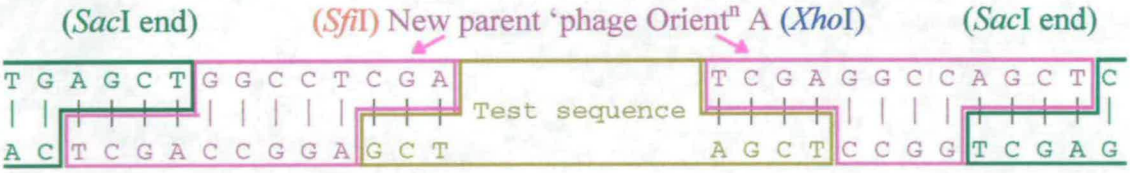
This was the design used. (The brackets surrounding the *Sac*I labels indicate that the site is destroyed by the innermost base-pairs.) This new parent 'phage that would be created by ligating this insert into DRL167 would differ from DRL167 in that a piece would have to be removed to ligate a new insert into it. It was not of course necessary to change so many base-pairs to remove the *Sfi*I site from the right side; the use of a sequence with 10 bp of asymmetry in the centre was however intended so as to reduce the tendency of the two constituent oligonucleotides to form hairpins rather than anneal with one-another. However, thinking about it again while writing this, instead of changing all of these base-pairs one could have deleted some of them; not

all however, otherwise a new *Sfi*I site would have been formed by bringing two GGCC motifs to the correct proximity.

This insert could be ligated into DRL167 in either orientation - as above, which will be referred to as 'Orientation A', or with the *Xho*I site to the left and the *Sfi*I site to the right, which will be called 'Orientation B'. Provided that these were not continually changing orientation by recombination, isolates with each orientation could be prepared as two new parent 'phage. New inserts testing sequences for hairpin-forming tendency would have a long and a short strand. With two parent 'phage it would only be necessary to make a single insert into the *Sfi*I and *Xho*I sites to test a sequence in both orientations rather than having to make, for instance, one insert with CAG on the short strand and CTG on the long strand and one with CTG on the short strand and CAG on the long strand.

ii) The inserts to the parent 'phage

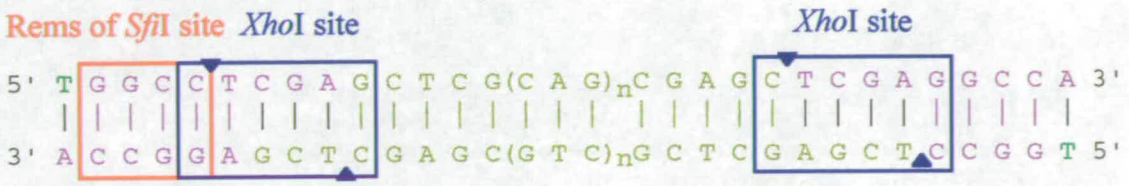
To restore symmetry, the minimum new insert would only be required to have the correct overhangs on the test sequence as shown below in olive green:



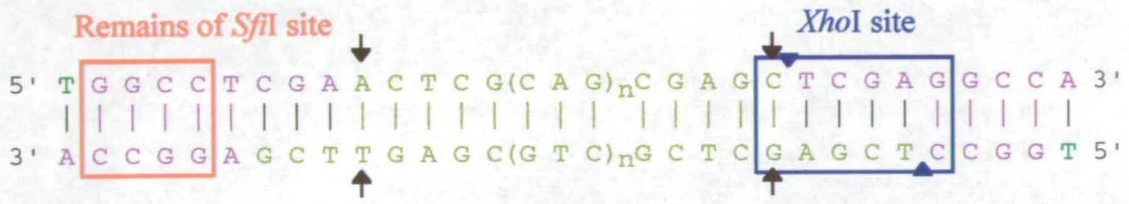
However, this might destroy both the *Sfi*I and the *Xho*I site. It would obviously be much cheaper to restore the *Xho*I site as this would only require a single correct base-pair on one side, and another base-pair on the opposite side to balance it. If this other base-pair was palindromically symmetrical then an *Xho*I site would be created on both sides. If it was not then only a single site would be recreated but there would be some asymmetry which might upset the results with the sequence under test. It was therefore decided to include a buffer sequence - symmetrically on either side of the test sequence - both to distance this 1 bp of asymmetry further from the axis of

symmetry of the palindrome [as it had been shown that base changes made less difference when further out (Davison & Leach, 1994a)] and to give the test sequence the same flanking context as in the inserts to DRL167 so that results could be compared. It was decided make this buffer 4 bp each side. Thus the plan of the new inserts was as below:

Symmetrical insert



Asymmetrical insert



again shown in Orientation A and using the same colour coding with the new insert in maroon and with $(CAG)_n{\cdot}(CTG)_n$ representing the test sequence. The asymmetric base-pairs are arrowed. The A·T pair was chosen to replace the G·C pair on the left because when the palindrome extruded these bases would form, across the hairpins, A-C and G-T mispairs which it was hoped might fit into the helix and cause less disruption than the other two possibilities. With either insert, an advantage of reforming one or two *Xho*I site(s) should be that one could test for the successful introduction of the insert by showing that ability to cleave with *Sfi*I had been lost but cleavability with *Xho*I retained.

iii) A ligation piece and primers for PCR and sequencing

Having plaque purified insert-bearing isolates form the original mixture of packaged parent 'phage, it would be necessary to investigate their orientations. This could be

done by cleaving the 'phage DNA with either *Sfi*I or *Xho*I and then conducting polymerase chain reactions on single palindrome arms which would not be subject to the problem of strong secondary structure formation. The central sequence of the insert could be used as a primer in conjunction with primers matching sequences outside the palindrome but this would not be useful with subsequent inserts to the new parent 'phage. It was planned to sequence the new asymmetric inserts using dichlororhodamine-labelled dideoxy terminators and DNA is usually amplified to sufficient concentration for this by PCR. Since the sequence of interest is right at the end of one of the *Xho*I fragments, a section of DNA for use as a primer-complement would have to be ligated on. This then could also be used for investigating the orientation of the original 'phage.

There were already primers for sequences outside the palindrome which had been chosen with the criteria of not annealing with one-another, having 3' ends that would not anneal elsewhere on the same oligonucleotide, and having the same annealing temperature, 70°C, on the simple formula T_m = (No. of G·C pairs × 4) + (No. of A·T pairs × 2). They were:

625J, Left arm	5'CCGTTGCAGA TGTTCTTGAA TACC
626J, Right arm	5'TTGGACTCAA GAATGCTGCC AGC

These correspond to bases 21,189 - 21,212 and 26,165 - 26,143 of the λ genome and are shown (in dark blue) in relationship to the palindrome of DRL167 in Appendix 1.

A new primer to pair with either of these was chosen by cutting out 23 small squares of paper and marking 12 of them with a C on one side and a G on the other and the other 11 with an A on one side and a T on the other, mixing them up in a cup and then drawing them out and laying them down in a line, then checking the result by the same criteria as above and searching the λ genome to check that there were no close matches. The sequence thus arrived at was:

*Xho*LigPri	5'GGGTAATCGT CATCAGTCTG TCG 3'

As this was to be the primer, the sequence to be ligated onto the cleaved 'phage was required to have a complementary sequence to this and to have a 3' end that would match the *Xho*I ends of the 'phage. By chance the last 3 bases on the 3' end of the primer were the same as the first 3 bases of the 4-base 5' overhang of *Xho*I (5'TCGA...) so the complementary piece only required to be 1 base longer, *i.e.*:

*Xho*Lig 3'CCCATTAGCA GTAGTCAGAC AGCT 5'

Since 'phage made with random central sequences would have to be constructed by ligating a single-strand degenerate oligonucleotide mixture to both the *Sfi*I and the *Xho*I ends of the cleaved parent 'phage, it was going to have to be possible to ligate a single strand to a double-strand 'sticky end' of the 'phage so it was not planned to anneal two oligonucleotides to make a double-stranded end-piece for PCR, hence the contentment to have the ligation oligonucleotide only one base longer than the primer, rather than four, as the two were not to be annealed prior to ligation. The oligonucleotide was of course 5'-phosphorylated.

As described below, the oligonucleotide was indeed ligated to the *Xho*I ends of the cleaved 'phage and used for the determination of the orientation of isolates by PCR product size, but PCR to the left side was not very strong and it was decided to try the effect of using a double-stranded end-piece so then a new oligonucleotide had to be ordered that was 3 bases shorter than the primer. This was then annealed to the last oligonucleotide above to make the following double-stranded end-piece (other way round from the above):

```
5' T C G A C A G A C T G A T G A C G A T T A C C C 3'
       | | | | | | | | | | | | | | | | | | | |
    3' G T C T G A C T A C T G C T A A T G G G 5'
```

(The new 20-base oligonucleotide is named '*Xho*Lig2'.) The ligation of this piece to the 'phage arms was clearly much more efficient than ligation of the single oligonucleotide as the PCR signals (using the same primers as before) were much stronger.

When the time came for sequencing, it was decided to use one of the computer programs available for primer selection to see whether better sites outside the palindrome could be found for use with *Xho*LigPri for amplification of palindrome arms prior to sequencing. The program used was 'Primer 3' (available on the Internet from http://www.genome.wi.mit.edu/cgi-bin/primer/primer3.cgi). The more sophisticated algorithm used by this program estimated that the T_m of *Xho*LigPri was not 70°C but 63·03°C and the primers selected for use with it were:

```
PalJDleft       AACCGAAGAA TGCGACACTG
PalJDright      GAACAACCTG ACCCAGCAAA
```

(which have estimated T_m of 61·24 and 61·08°C respectively). They correspond to bases 21,054 - 21,073 and the complementary sequence of 26,205 - 26,186 of the λ genome and are shown on the sequence in Appendix 1 (in green). After amplification of individual palindrome arms, sequencing of their central ends was initially tried with a primer within the palindrome which gave products of ~100 nt. This should have been a good strategy for manual sequencing but the automated sequencer had difficulty in recognizing such short strands so the above primers (PalJDleft and PalJDright) were used for sequencing of the PCR products. Sequencing was of course done towards the centre of the palindrome, and not from the centre outwards, because the crucial sequence was at the centre.

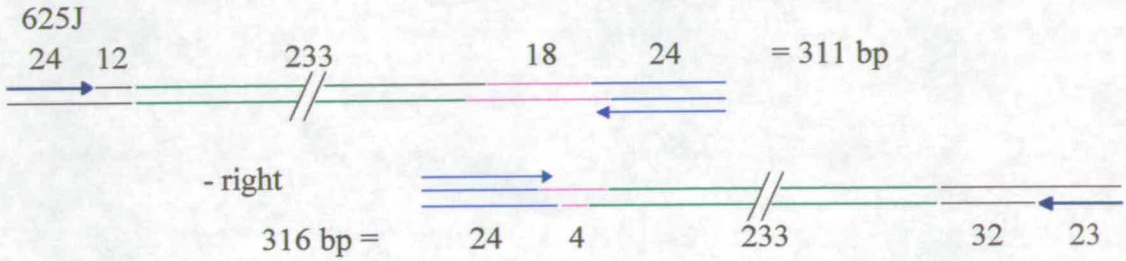
Construction and selection

The new parent 'phage were constructed and isolates plaque-purified as before but of course were tested for the presence of the insert by the presence of the *Sfi*I and *Xho*I sites (rather than the *Bsa*I site) as well as by the absence of the *Sac*I site. The next job was to find isolates in which the insert had ligated into the *Sac*I site in each of the two orientations. Four isolates were plaque-purified as this would give only a 1 in 8 chance that all would have the same orientation (1 in 16 chance of all in orientation A and 1 in 16 of all in orientation B). Plate lysates and DNA

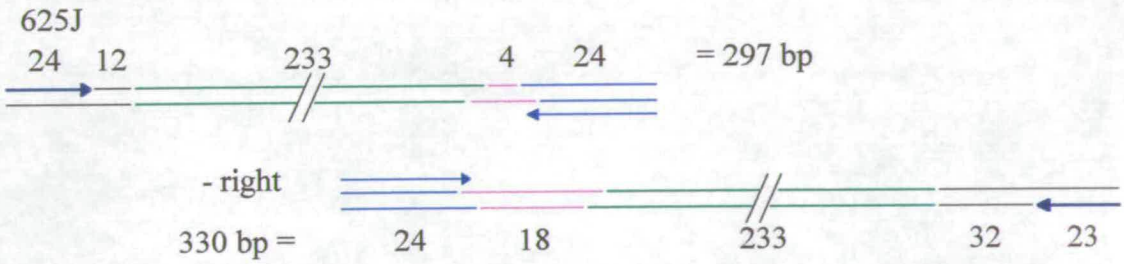
minipreps were prepared for each isolate and then the DNA was cleaved with *Xho*I and ligated to, initially, the single oligonucleotide *Xho*Lig and later the double-stranded end-piece, and PCR was carried out with on portions of the ligation products with *Xho*LigPri paired with each of the primers 625J and 626J.

The sizes of the expected products, with lengths measured along the strand labelled, were as shown:

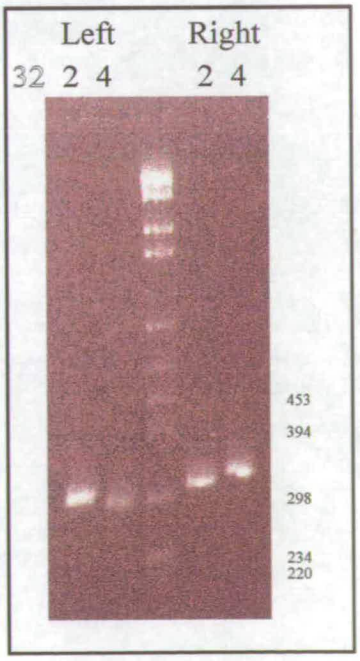
Orientation A - left



Orientation B - left



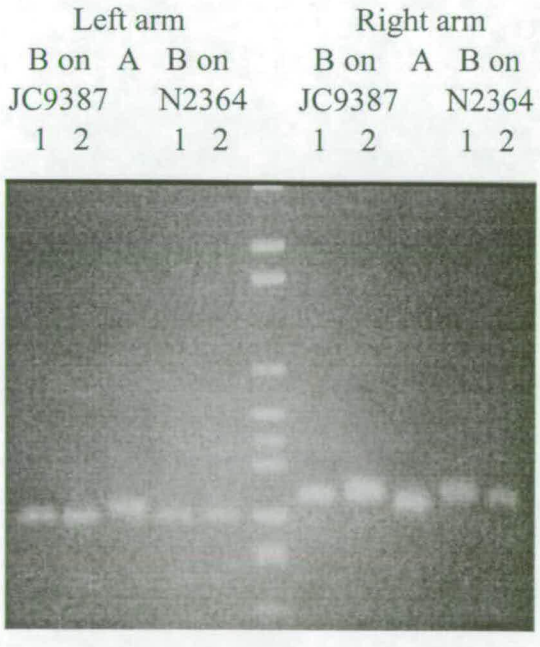
Thus the left arm products are smaller than the right and orientation changes the size by 14 bp. This was construct 32 and isolate 2 was found to have Orientation A and isolates 1, 3 and 4 had Orientation B. The picture on the right shows PCR products of the left and right arms of 32,2 and 32,4 on a 3% 'Nuseive 3:1' agarose/TBE gel. The central lane is Boehringer-Mannheim Marker VI and nearby band sizes are marked. These 'phage have since been given lab. names DRL257 (for Orientation A) and DRL258 (for B).



'The flipping experiment'

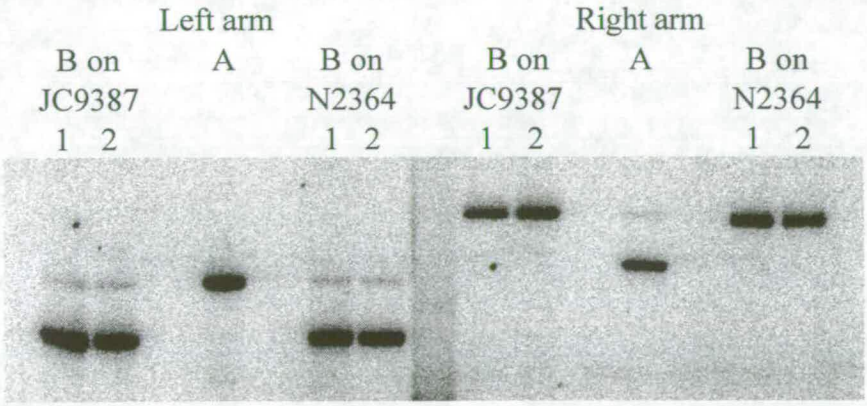
This was an investigation of the possibility that the palindrome centre might change orientation by recombination between the palindrome arms frequently enough in the bacterial strain used for the plaque assay, N2364 (*sbcC, rec*$^+$) to make investigation of the effect of orientation pointless.

Lysate of 32,4 (the orientation B 'phage) was diluted to 10^{-6} and 1 µl portions of this were added to each of two portions of 250 µl JC9387 (*recBC, sbcBC*) plating cells (used for plaque purification, preparation of lysates and titring, see Chapter 2) and two portions of 250 µl N2364 plating cells, allowed the usual 15 - 20 min to adsorb, and then each culture was plated in the usual way (Chapter 2) with 2·5 ml of BBL top agar on a BBL plate and incubated overnight at 37°C. Next day, one plaque was taken from each plate and plate lysates were prepared in the usual way. DNA minipreps were then made from each of the lysates, the DNA was cleaved with *Xho*I, the small DNA pieces for PCR were ligated on and PCR was carried out on each palindrome arm in each preparation. Below is shown the result of electrophoresis of the products on a 2% agarose gel. The marker DNA (central lane) is the same as before.



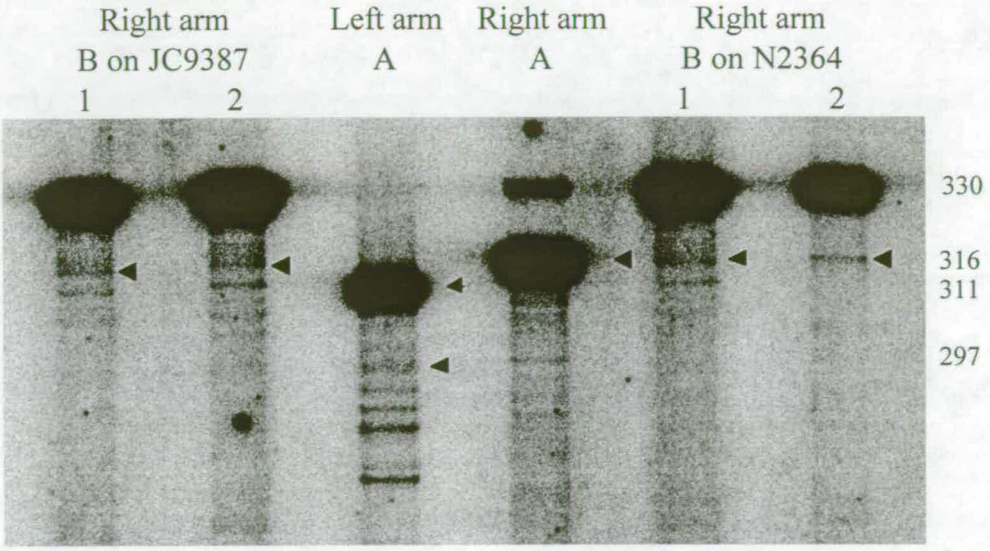
It can be seen that from all four preparations the left arm PCR products were the same size and were smaller than the left arm product of the orientation A 'phage, included for reference. Likewise the right arm products of all four preparations were the same size and larger than the right arm product of the orientation A 'phage. The number of cycles of 'phage replication in the cultures is

not known but clearly there has not been a reversal of orientation of the palindrome centre at an early stage in any of the preparations. On this gel there does not appear to be any evidence that reversal has occurred at a later stage giving rise to minor bands due to 'phage with the opposite orientation. Several further gels were run without changing this impression and it was decided that flipping of the palindrome centre by recombination was not going to be a serious problem and investigations comparing the effect of inserting sequences in either orientation, to be described below, were carried out. However, later, DNA from the PCR reactions was 5'-end-labelled with ^{32}PγdATP and run on polyacrylamide gels in order to try to detect and quantify any flipping that might have occurred. Below is shown the relevant section of a PhosphorImage of a 5% denaturing gel with the samples loaded in the same order as on the previous gel. (Marker DNA was included in the central lane but did not label properly.)



It can be seen that the PCR products of the left palindrome arm of all four cultures of the Orientation B 'phage show a minor band with the same mobility as the left arm product of the Orientation A 'phage, suggesting that flipping of the palindrome centre has occurred in some 'phage. However, if flipping has occurred then PCR of the right arms should show minor bands with the mobility of the right arm of the Orientation A 'phage and such bands are not seen. Strangely also, PCR of the right arm of the A orientation 'phage, which was not recultured since its original preparation, shows a minor band with the mobility of the right arm of the B orientation 'phage but this is not matched by a minor band in the left arm PCR.

A possible explanation might be that some flipping has occurred in all cases and that PCR is more successful when the sequence contains the asymmetric insert, *i.e.* the left arm in Orientation A and the right arm in Orientation B. To check for the presence of faint bands to support this, portions of all five labelled right arm products and of the orientation A left arm product were subjected to electrophoresis on another 5% denaturing polyacrylamide gel and exposed to a PhosphorImager screen for 16 days. The result is shown below.



It can now be seen that there are faint bands of mobility 316 bp that could represent flipping. The ones relating to the cultures on N2364 are a little clearer and image quantification showed No. 2 to contain ~ 1·2% of the amount of DNA in the 330 bp band. For the A orientation 'phage (not recultured) the amount of DNA in the 297 bp band of the left arm is ~ 1·6% of that in the 311 bp band. These are in contrast to the amount of DNA in the 330 bp band of the right arm of the A 'phage, which is ~ 6·5% of that in the 316 bp band. As the long exposure has revealed various other faint bands in all lanes one cannot be completely certain that the bands of the correct sizes do represent flipping but it seems a reasonable assumption that they do. Quantitative PCR is difficult and these results cannot be taken to indicate the real frequency of flipping. They do, however, suggest that it is not enough to prevent investigation of whether orientation affects plaque area.

Experiments with the new 'phage

By the time of completion of the review (Darlow & Leach, 1998a), which forms the basis of Chapter 5, the time and grant for this project had expired, but it seemed a pity not to include any use of the new 'phage having constructed it. However, it was not possible to continue the work until every peculiarity of the results had been explained and/or corrected. Therefore the results to be presented are incomplete and I shall present them as they unfolded, rather than as a body of work from which some results are missing.

Investigation of the effects of orientation and asymmetry

Before testing any new sequences for hairpin-forming tendency it was felt necessary to try sequences already tested in DRL167 for comparison. The sequences selected were $d[(CAG)\cdot(CTG)]_{1-3}$ which, as we have seen, give a strong √ pattern of plaque sizes. These could be used both for testing the effect of orientation and that of the one base-pair asymmetry of the insert described earlier. Symmetric and asymmetric inserts were therefore constructed with each of these three sequences in the centre.

It was originally decided to ligate both types of insert into the A form of the new 'phage and just the symmetrical ones into the B form (and maxipreps of DNA of both forms were made). The effect of orientation would then be seen by comparing the plaque sizes of 'phage made with symmetrical inserts in the A and B parent 'phage, and symmetric and asymmetric inserts would be compared in the A parent. However, after encountering difficulty in cloning with the A parent it was found that the prepared DNA had become degraded but that that of the B parent had not. By this time isolates which appeared from restriction digests to be correct, had been made from the A parent with the $(CAG)_3$ symmetrical insert and all three

asymmetrical inserts so, rather than prepare more parent 'phage DNA, it was decided to ligate both symmetrical and asymmetrical inserts into the B parent.

Having tested the remaining constructs by restriction digestion, plaque assays were carried out. As described in Chapter 3, six is the largest number of 'phage that could practically be compared in one assay with enough plates for each 'phage to give reasonable accuracy, so in the first assay isolates of 'phage made from the orientation A parent with the symmetrical insert containing $(CAG)_3$ and the asymmetrical inserts containing $(CAG)_{1\text{-}3}$, and isolates from construction with the orientation B parent and the symmetrical insert containing $(CAG)_1$ were compared with the reference 'phage, DRL176, as the sixth. The results are shown in Figure 7.1.

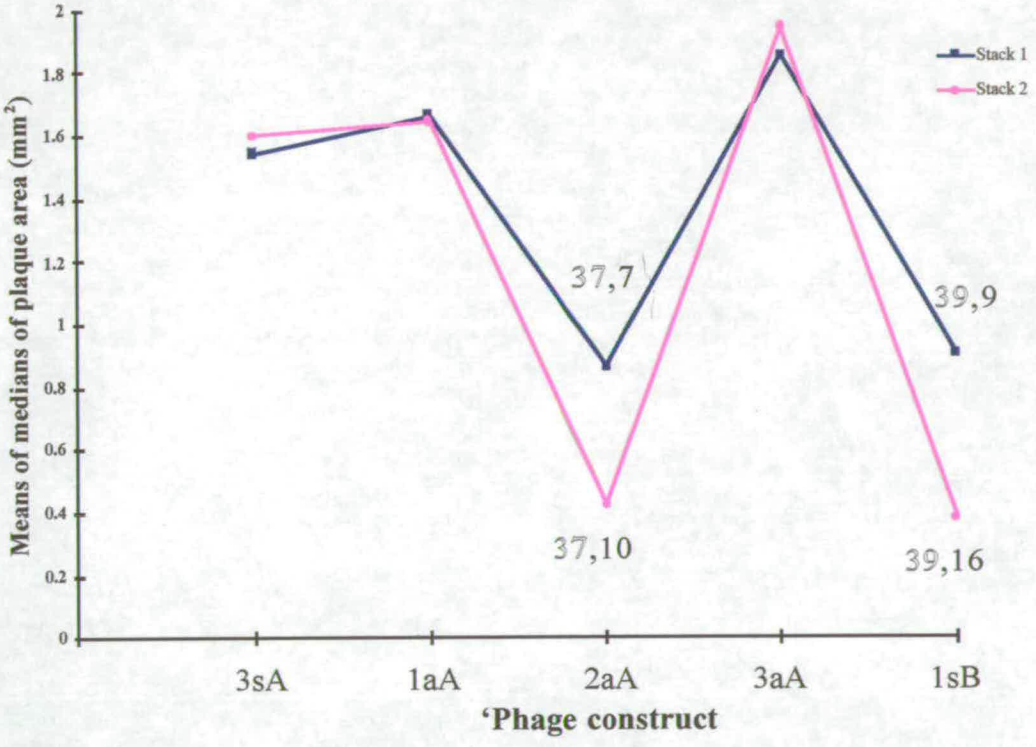


Figure 7.1 Results of the first plaque area assay with 'phage constructed with the new parent 'phage. The lines connect results from 'phage grown on plates from the same stack. On the X axis the shorthand names are used as in Table 2.2. On the chart are marked the names of individual isolates which gave different results.

As usual, four plates were used from each of two stacks for each 'phage and where there were two positive isolates for a 'phage, one was grown on the plates

from one stack and the other on those from the other. It can be seen that for both 2aA [$(CAG)_2$, asymmetric insert, orientation A] and 1sB [$(CAG)_1$, symmetrical insert, orientation B], the two different isolates from the same construct gave plaques of different sizes though the two results for each of the other 'phage were very similar. Because the results of 37,7 and 39,9 were very similar, as were those of 37,10 and 39,16, it was thought that possibly material could have been placed in the wrong tubes at some stage of construction. If so, which was which? Since a 'phage with a palindrome containing the central sequence $d(CAG)_2$ had previously been found to produce the smallest plaques, it was thought that perhaps 37,7 was really an isolate of 1sB and that 39,16 was really an isolate of 2aA. Rearranging the results according to this scheme gave the picture shown in Figure 7.2.

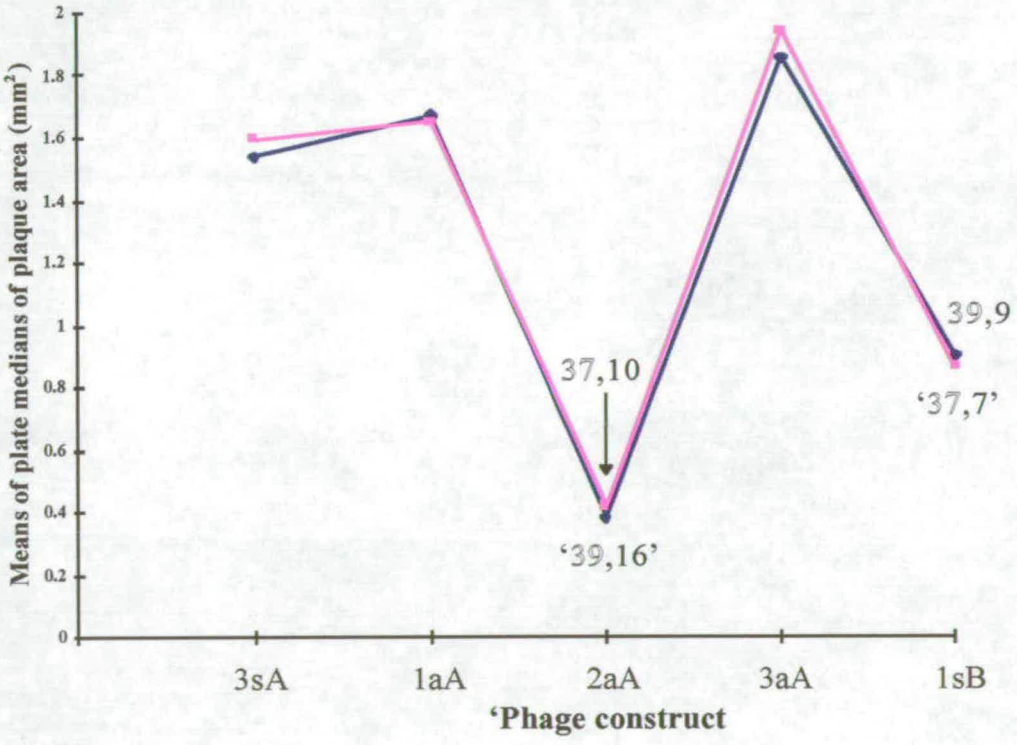


Figure 7.2 Results from Figure 7.1 with 37,7 and 39,16 exchanged.

This appeared to give very tightly grouped results for every 'phage. The 'phage remaining to be tested were 2sB, 3sB, 1aB, 2aB and 3aB. It would be possible to test these in the same assay with the usual reference 'phage, DRL176, for

scaling of results from the two assays so that they could be considered together, but it was important to see whether there was the same √ shape for 'phage with inserts with $(CAG)_{1-3}$ whatever the conditions of orientation, symmetry or 'phage context (DRL167 or the new parent 'phage). Therefore it was considered better to compare the 1-3sB and 1-3aB 'phage in the same assay without DRL176 and to use the results of 1sB for scaling, as their plaques would have been measured in both assays, and to take the risk that 37,7 and 39,9 might not really both be 1sB. The main results of this assay are shown in Figure 7.3.

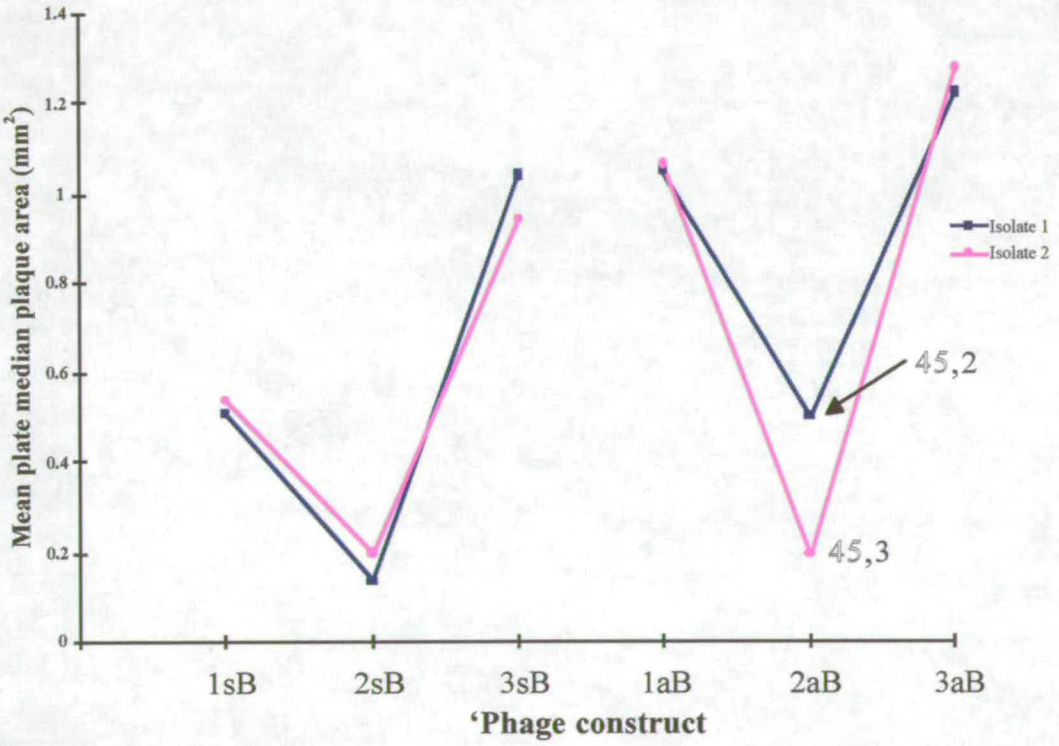


Figure 7.3 Plaque assay comparing isolates of 'phage with symmetrical and asymmetrical inserts in orientation B. Names of individual isolates are only given where they give substantially different results.

In addition to the results shown, three other 'phage were each plated on one spare plate from the same stacks. For 'phage construct 39 (supposed to be 1sB) three plaque-purified isolates had all been found to contain an *Xho*I site and to lack an *Sfi*I site and had therefore all appeared to contain the insert. The one not so far tested

for plaque size was 39,25 and this was plated along with one plate each of 39,16 and 37,10 (that were both now thought to be isolates of 2aA). The result of these was that the plaque size of 39,25 was about the same as that of 39,9, tending to confirm my suspicion that 39,9, not 39,16, was the correct 1sB isolate. The plaques of 39,16 and 37,10 were again smaller and about the same size as each other.

Three main findings are immediately apparent from the results: (i) the shape of the $(CAG)_{1\text{-}3}$sB plot is similar to that for the plot of the plaque sizes of 'phage with $(CAG)_{1\text{-}3}$ inserts in DRL167 (which were also symmetrical) but (ii) the 1-3aB shows almost a V shape rather than a √ shape because the $(CAG)_1$ plaque size is much larger than that of $(CAG)_2$ and almost as large as that of the $(CAG)_3$ plaque size. This is just the same pattern as was seen with 1-3aA (Figures 7.1, 7.2) and tends to confirm the earlier result and to suggest that a 1 bp asymmetry at the margin of the insert might have a much larger effect on the plaque size when the central (trinucleotide) sequence is short than when it is longer. (iii) With the 2aB isolates we again see the phenomenon of two different isolates making different-sized plaques. The 45,3 result was about the same as that of the two tested isolates of 2sB, 40,61 and 40,73, so it seemed likely that the 45,2 result was the correct one.

Three isolates of each 'phage construct (2sB and 2aB), that had been selected as still containing long palindromes, had been plaque purified, but one of the isolates of 2aB had been found by restriction digestion to lack the insert. The three isolates of the 2sB construct all appeared to contain the insert. Therefore, if 45,3 was really an isolate of 2sB, the untested isolate supposed to be 2sB (40,59) should really be an isolate of 2aB and should produce a plaque size the same as that of 45,2. All the results described were apparent on inspection of the plates on removal from the incubator before any image-quantification had been carried out. Accordingly, the untested isolate, 40,59 was plated on a spare plate along with one more plate of 45,2 and incubated. However, the plaques of 40,59 were tiny, just like those of 40,61

and 40,73, so this time the finding of two different plaque sizes from isolates of the same construct (45, 2aB) could not be explained by mixing up tubes. It was clearly necessary to start checking sequences.

As described earlier, after cleavage with *Xho*I and ligation of the end-piece for PCR, each palindrome arm could be amplified and sequenced. If the insert was asymmetrical, cleavage would occur on only one side of the insert, so not only could it be sequenced but it could be seen to which arm it was still attached: the left arm if the orientation was A, the right if B. If the insert was symmetrical, it would be cleaved at both sides and lost but the palindrome size could be compared with the sizes of palindromes that could be sequenced and the palindrome arms could still be sequenced if necessary. The following findings emerged:

Short name	Isolate	Results
3sA	35,4	Palindrome has correct length (same length as 38,12).
1aA	36,8	$(CAG)_1$ insert on left arm, *i.e.* 1aA
	36,14	ditto
2aA	37,7	Is 2aA, contrary to what was expected.
	37,10	Palindrome centre absent from both arms (*i.e.* insert has become symmetrical). Size of fragment shows deletion.
3aA	38,12	Is 3aA.
1sB	39,9	Palindrome symmetrical, as expected. Size correct.
	39,16	Palindrome symmetrical. Centre deleted.
2sB	40,59	Centre deleted.
	40,61	Centre deleted.
	40,73	Two different palindrome sizes, both too large
3sB	41,29	Palindrome has correct length (same length as 38,12).
1aB	44,1	Is 1aB.
2aB	45,2	Is 2aB.
	45,3	Symmetrical
3aB	46,4	Is 3aB.

The first thing to notice is that isolates of constructs 37 and 39 had not been mixed up. The isolates that made the larger plaques, 37,7 and 39,9 (see Figure 7.1) were both what they were supposed to be; the plaques of a 'phage with two d(CAG) trinucleotides at the palindrome centre and a 1 bp asymmetry between the arms (and orientation A) had made plaques about the same size as one with only one d(CAG) trinucleotide whose palindrome arms had perfect symmetry (and orientation B). The plaques of 37,10 and 39,16 were the same size because the 'phage were identical because they had both suffered the same mutation. After cleavage, both had identical palindrome arms bearing no insert. The palindromes had become symmetrical but this had not been by correction of the 1 bp asymmetry but by deletion of the central sequence (which had been different in the two inserts). This could only have been between the two direct repeats provided by the altered and correct *Xho*I sites to make a single correct central *Xho*I site (Figure 7.4, overleaf) because deletion between any other points would not have resulted in identical arms with *Xho*I sites at their central ends.

This is a little surprising because deletion within palindromes usually results in asymmetry. (However, this thesis is already more than long enough and I am not going to discuss this here.) The same problem did not arise with the inserts to DRL167 because the restriction site of *Bsa*I is asymmetric and was placed in opposite orientations on the two sides so did not provide a direct repeat. However, *Bsa*I could not be used in this case because it cleaves the λ genome at two other places, and other available enzymes with asymmetric recognition sites cleave at more sites.

The same mutation had evidently occurred to 45,3 but again, luckily, the other isolate had the correct sequence. Unfortunately with construct 40, 2sB, one of the isolates used in the plaque assay, 40,61, was found to have the deletion and the other, 40,73, bizarrely showed two palindrome bands, both larger than intended (by about 20 and 25 bp). There was no time to explore what mutations might have

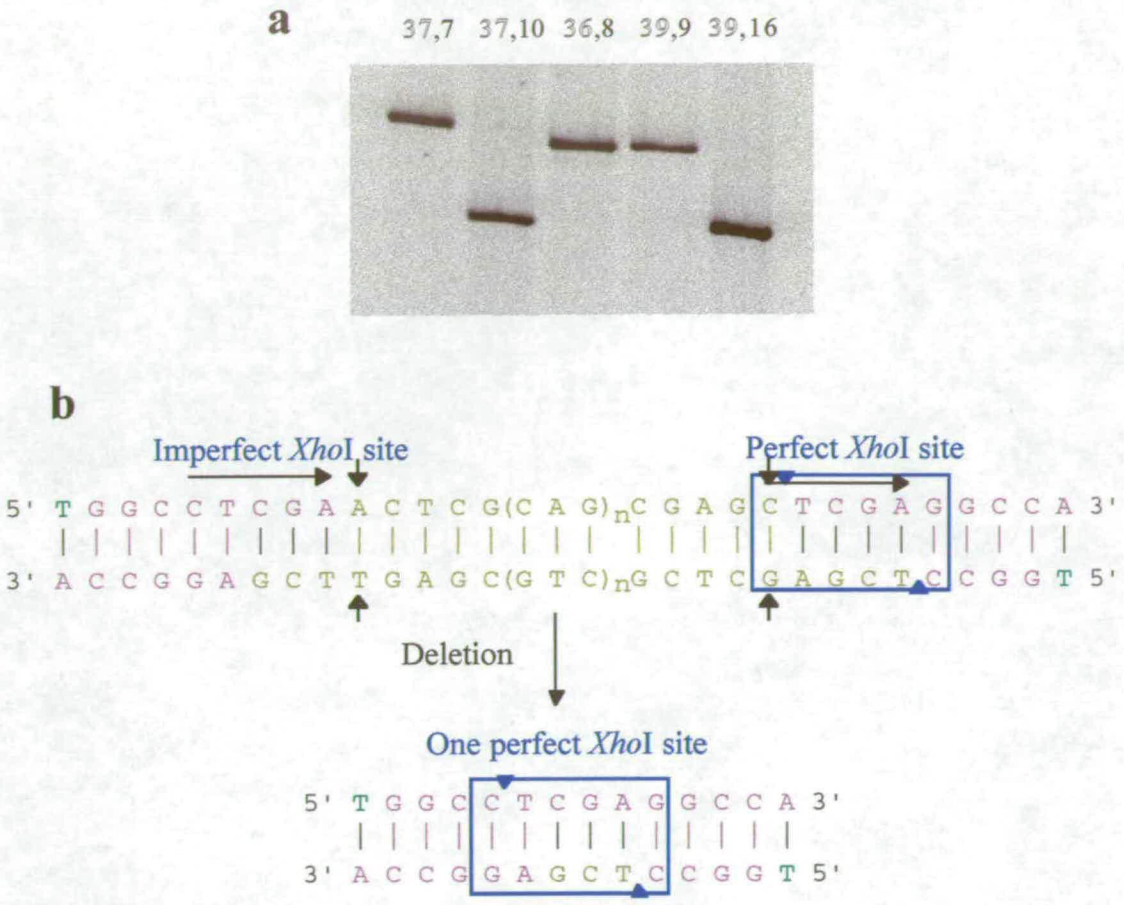


Figure 7.4 Loss of palindrome centres by deletion between direct repeats. (a) PhosphorImage of end-labelled palindromes run on a 5% denaturing polyacrylamide gel. The palindromes of 36,8 and 37,7 had been shown by sequencing to contain 1 and 2 central d(CAG) trinucleotides respectively so act as size markers. It can be seen that the palindrome of 39,9 is the correct size but that those of 37,10 and 39,16 are deleted and of the same size. (b) What the palindrome central sequence should be, showing 5 bp direct repeats, and how a (14+3n) bp deletion results in a perfect palindrome with a single perfect *Xho*I site at the centre.

---

caused this. A spare isolate, 40,59, had been plaque-purified and checked by restriction digestion with *Sfi*I and *Xho*I but not used for the assay so its palindrome was checked by electrophoresis. Unfortunately it too proved to have the deletion. I could not afford to go on any longer to try to produce a correct construct of 2sB and carry out a further plaque assay to compare it with the other 'phage so had to be

content with an incomplete set of results. The results of the two plaque assays of Figures 7.1 and 7.3 could be combined. This is shown in Figure 7.5. Though 37,7 had proved to be 2aA not 1sB, it had been plated in both assays so its plaque areas in the two assays could still be used for scaling the other results. The plaque sizes of both 37,7 and 39,9 were therefore used for scaling. The mean of their plate medians of plaque area in the first assay was 0·88 mm^2 and in the second 0·52 mm^2 so the results of the second assay were multiplied by 0·88/0·52 = 1·69.

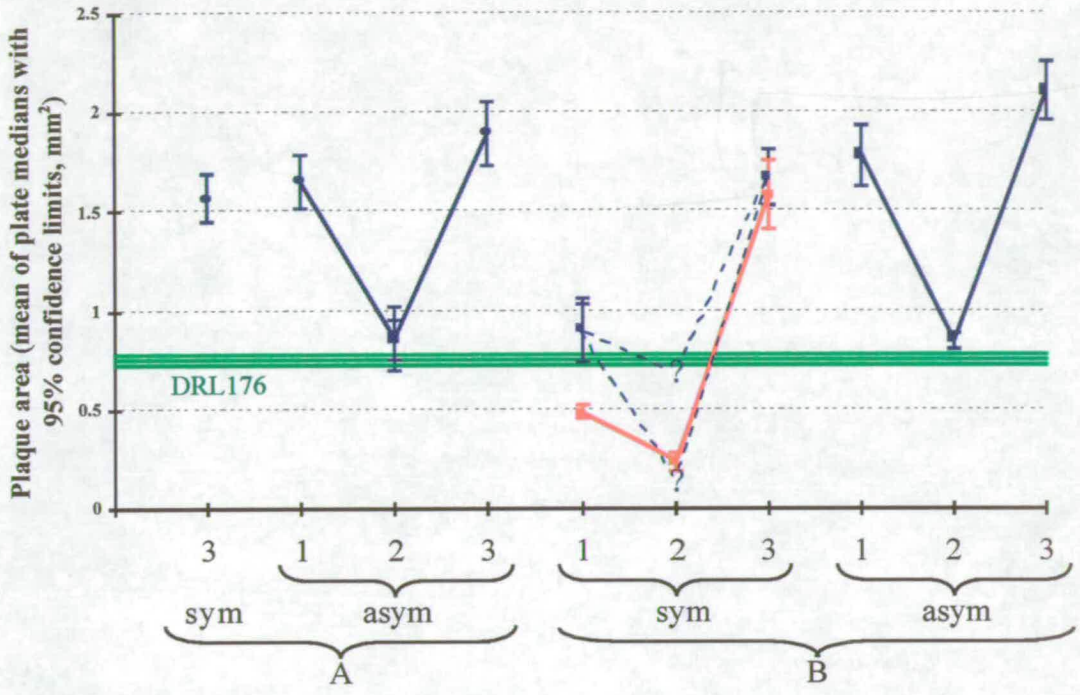


Figure 7.5 Combined results of plaque assays shown in Figures 7.1 and 7.3 showing only results from 'phage shown to have correct structures. There are two error bars on the results of the 2aA and 1sB 'phage since they were included in both assays. DRL176 was only included in the first assay but its result was used to place the results obtained with $d(CAG)_{1\text{-}3}$ inserts in DRL167, shown superimposed in red. The original results of the latter were multiplied by 0·74 for this comparison.

From this chart it can be seen that: 1. The results for the asymmetric 'phage are very similar in the two orientations. The $(CAG)_2$ results are essentially the same. The $(CAG)_1$ and $(CAG)_3$ plaque areas are apparently slightly larger in orientation B than in orientation A but they were not measured at the same time and the 95%

confidence limits overlap. This is explored further below. The one sequence in symmetrical inserts tested in both orientations, $(CAG)_3$ also shows virtually the same result in both orientations. These results tend to confirm the suspicion that orientation of the central sequence in the palindrome would not affect plaque size, mediated, we believe, by extrusion of the palindrome to a cruciform. 2. The results for the asymmetric 'phage are larger than those for the symmetrical 'phage. That for 1sB is less than half that for 1aB. The result for 3sB is substantially smaller than that for 3aB, the 95% confidence limits not overlapping. Though we have no result for 2sB we can assume that it would also be smaller than that of 2aB. 3. In both orientations, the result from $(CAG)_1$ with the asymmetric 'phage is much closer to that from $(CAG)_3$ than was the case with the original 'phage (shown in red) but the results confirm that $d[(CAG)\cdot(CTG)]_2$ is a strong hairpin-nucleating sequence and therefore the asymmetric 'phage, in which the inserts can be checked by sequencing, could be suitable for checking the hairpin-forming tendencies of other sequences. 4. The plaque size obtained for 3sB is similar to the scaled result for $d[(CAG)\cdot(CTG)]_3$ in the centre of the original 'phage but the result of 1sB appears to be about 80% larger than its counterpart in the original 'phage. One would not be surprised if the plaque sizes were not identical since the flanking sequences beyond the buffer of 4 bp either side, are not identical, but this seems rather a large difference. It would be necessary to include the 'phage in the same assay to check this.

Without a result for 2sB it cannot be said whether the shape of the 1-2-3 line is the same in the old and new 'phage but, as I have said, investigations could not be continued indefinitely. This is not, however, the end of the results, because I had already performed some other investigations before discovering that all the prepared isolates of 2sB were defective.

In order to be more certain about a lack of effect of orientation on plaque size, the 1-3aA and 1-3aB 'phage were compared in the same plaque assay. Because it has been seen that plaques on plates from further down in the same stack tend to be a little larger, a precaution was taken to try to ensure that this did not lead to a false

conclusion that one orientation produced larger plaques than the other. As usual, four plates were dealt out to each 'phage from each of two stacks, but from the first stack they were dealt in the order 1aA, 1aB, 2aA, 2aB, 3aA, 3aB, and from the other 1aB, 1aA, 2aB, 2aA, 3aB, 3aA. The results are shown in Figure 7.6.

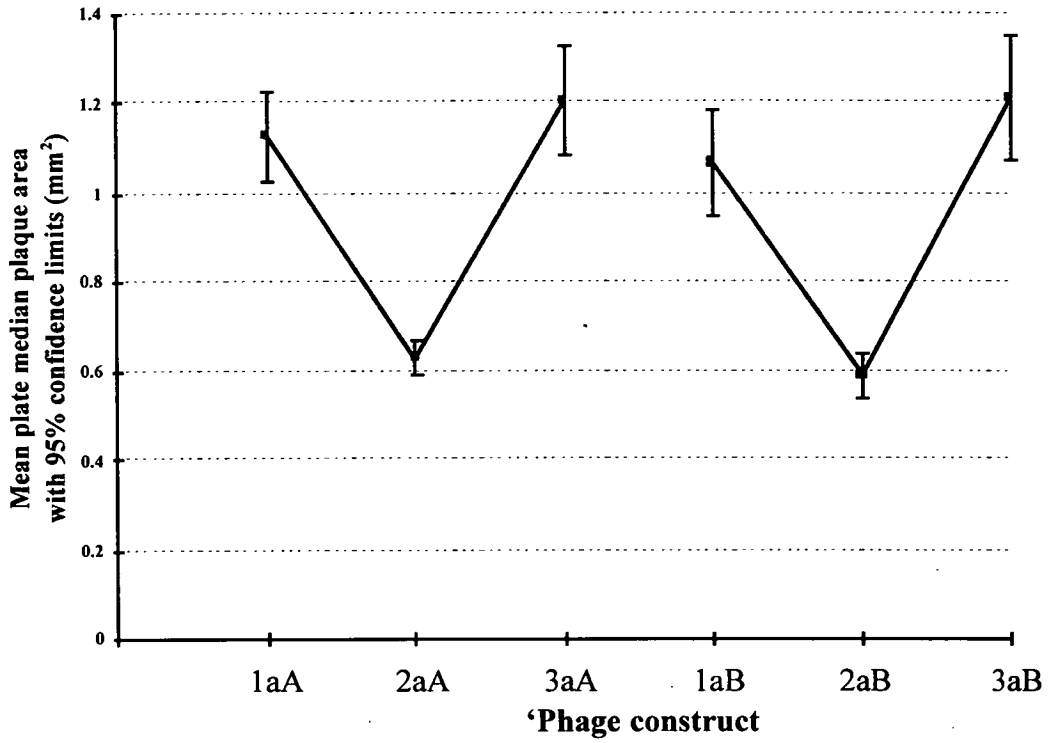


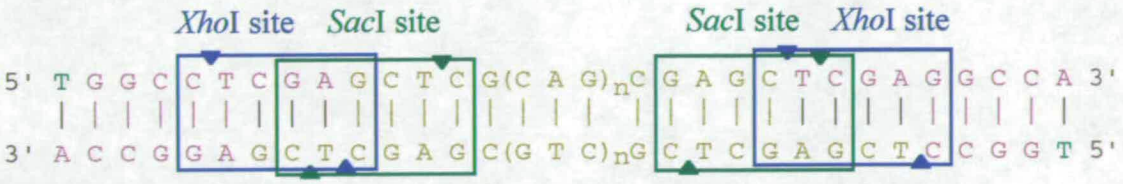
Figure 7.6 Plaque assay comparing isolates of 'phage with asymmetrical inserts in both orientations.

These results show that plaque size is indeed not affected by the orientation of the insert into the palindrome. I tried labelling the data of the A 'phage as B and the B 'phage as A in stack 2 to plot the data are as though the plates of any 'phage came from the same positions in each stack. This had the effect of bringing the lines plotted for each of the separate stacks closer together but because there was quite a wide plate-to-plate variation for any 'phage within each stack, there was little effect upon the 95% confidence intervals.

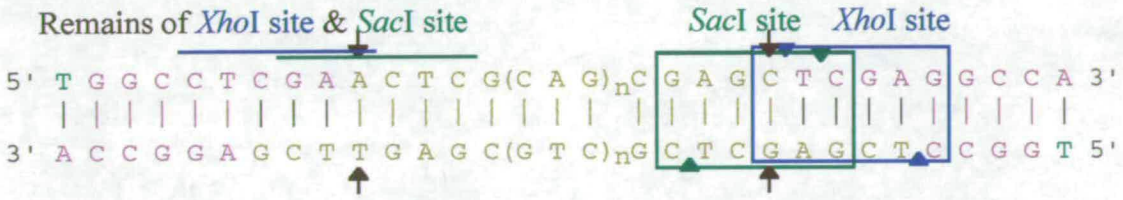
To conclude this section, I shall mention a useful point, for anyone who might continue this work, that I only noticed whilst drawing the diagrams for this thesis. The combination of using, in the new 'phage construct, the recognition sequence for

*Xho*I, d(CTCGAG), with the same sequence that flanked the test sequence in the usual DRL167 inserts, d(CTCGtest sequenceCGAG), has resulted in the inadvertent inclusion in the inserts to the new 'phage of *Sac*I sites, d(GAGCTC), two in the symmetrical inserts and one in the asymmetrical ones, as shown below.

Symmetrical insert



Asymmetrical insert



In the previous system, using DRL167, the insert was ligated into a *Sac*I site. It destroyed the site but introduced *Bsa*I sites. Identification of 'phage which contained an insert could thus be achieved not only by detecting lack of a *Sac*I site but presence of a new cleavage site for *Bsa*I. The new parent 'phage (of either orientation) has an *Sfi*I site and an *Xho*I site and after replacement of the DNA section between these sites with a new insert, success has been detected by loss of the *Sfi*I site while the *Xho*I site remains. Thus there has only been a negative test for insertion: loss of the *Sfi*I site. The *Sac*I site in the insert can be used as a positive test but more importantly, it will detect 'phage which have suffered *Xho*I→*Xho*I deletion. These 'phage have lost the *Sfi*I site and retain the *Xho*I site, just like intact 'phage with inserts, but the *Sac*I site will be missing. The presence of two *Sac*I sites also provides direct repeats between which deletion might possibly occur, and this test would not, of course, detect that, but I have not so far witnessed such a deletion.

A trial of d(GAA)·d(TTC) repeats

As mentioned in Chapter 1, d(GAA)·d(TTC) repeats were found to be highly unstable at one locus in *Gallus domesticus* (Epplen *et al.*, 1991), to be present in expanded tracts in some humans (Lindblad *et al.*, 1994), and to be the cause of Friedreich's ataxia when they expand in an intron of the frataxin gene (Campuzano *et al.*, 1996). The d(GAA) strand was found to form secondary structure, thought to be a quadruplex (Lee *et al.*, 1980; 1990) and the two strands were shown to be able to form a triplex (Shimizu *et al.*, 1989; Hanvey *et al.*, 1989), as they have again more recently (Gacy *et al.*, 1998; Mäueler *et al.*, 1998; Mariappan *et al.*, 1999), but from theoretical considerations the single strands were not expected to form hairpins (Mitas *et al.*, 1995a) or even to form any secondary structure at all (Gacy *et al.*, 1995).

Since the plaque assay I have used depends upon palindrome extrusion, it is really only suitable for determination of hairpin-forming tendency (which is not to say that triplexes might not cause problems for the replication of bacteriophage and thereby, small plaques). There were, however, two reasons why we considered that it might be worth testing this repeat sequence in the same way as the others. Firstly, it seemed a good idea to have results from a sequence not expected to form hairpins to compare with those that were. Secondly, it has been shown that 5' G·A 3' is a very good loop-closing pair and can form d(GNA) loops with a single unpaired base (Zhu *et al.*, 1995; Yoshizawa *et al.*, 1997; van Dongen *et al.*, 1997; Réfrégiers *et al.*, 1997; Jollès *et al.*, 1997) and therefore a single d(GAA) triplet at the centre of the palindrome might produce very small plaques though this effect would be expected to be diluted rapidly with addition of more copies if they did not form stable hairpins.

The purpose of testing the new 'phage with asymmetric inserts (that could be sequenced by virtue of their asymmetry), with a previously-tested series of sequences, was to see whether they would be suitable for testing the folding potentials of as-yet-untried sequences. The $d[(CAG)\cdot(CTG)]_{1\text{-}3}$ pattern with

asymmetric inserts to the new 'phage was similar, though not identical to the pattern obtained with symmetrical inserts to DRL167 and it was decided to go ahead and try d(GAA)·d(TTC) repeats with asymmetric inserts in the new 'phage. The orientation B parent 'phage (DRL258) was used (simply because the DNA preparation was of higher quality) and the inserts had $d(TTC)_n$ on the short strand, $d(GAA)_n$ on the long, so $d(GAA)_n$ was inserted into the top strand of the 'phage in this orientation (not that this should make any difference). There was not time to test the repeats in more than one frame and naturally the frame d(GAA)·d(TTC) was chosen because this was the one that might form a very tight loop.

As in the earlier work, a series of five 'phage were constructed, with $d[(GAA)\cdot(TTC)]_{1\text{-}5}$. and these were assayed in the usual way with DRL176 as the reference 'phage. The raw results are shown in Figure 7.7.

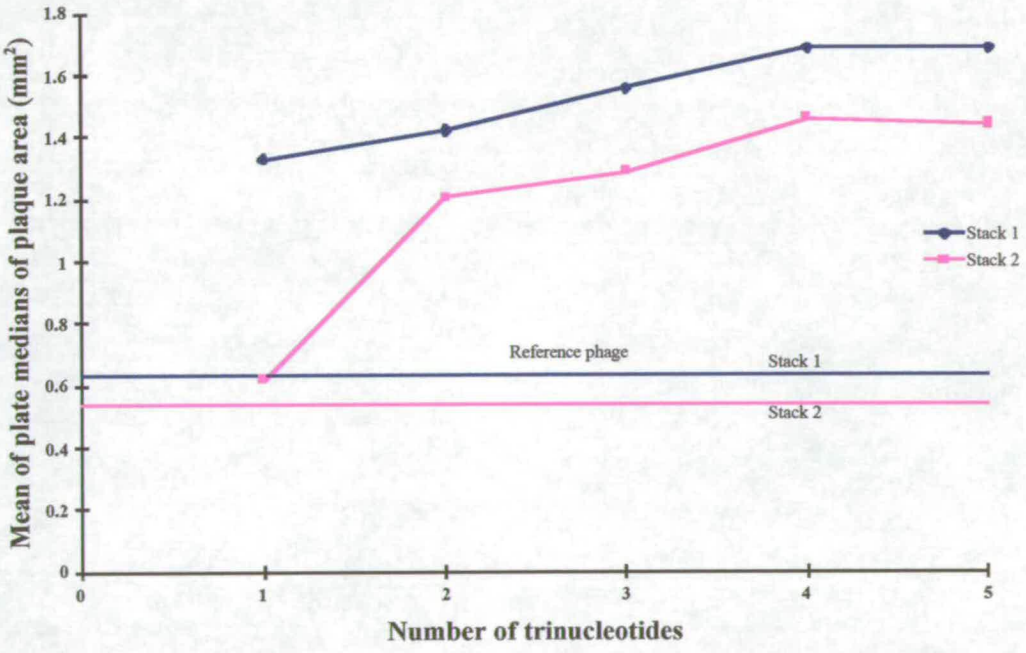


Figure 7.7 Raw results from the assay of d[(GAA)·(TTC)] repeats, results from plates from each stack shown separately.

In this assay there was a large difference in the results from the two stacks but for 'phage with 2 - 5 repeats the ratio between the mean of plate medians from the two stacks was almost exactly the same as the ratio between the respective

results of the reference 'phage and when the results from the d[(GAA)·(TTC)]-containing 'phage from stack 2 were multiplied by the mean of medians ratio for the reference 'phage, stack 1/stack 2, = 1·175, the lines of the d[(GAA)·(TTC)]-containing 'phage were overlain. The exception, of course, was for $d[(GAA)\cdot(TTC)]_1$ where there is a very obvious difference in the plaque sizes formed by the isolates used in the two stacks, 47,1 and 47,3, that was not corrected by this adjustment.

Sequencing was carried out, by the previously described method, of the right palindrome arm of both of the isolates with one copy of the trinucleotide and one isolate each for the others, some being isolates used in the first stack and some used in the second. In the other examples in this chapter in which two isolates from the same construct gave different results, it was found that the one forming the smaller plaques was the aberrant one, having become symmetrical by deletion. However, in this case the one forming the smaller plaques, 47,3, was found to have the correct sequence. 47,1, that was used on plates from stack 1, was found to have formed larger plaques because it had become more asymmetrical by means of a single base-pair deletion, the 4th 3' to the central trinucleotide. The sequences in the relevant region, in the direction of outside to centre of the palindrome, 5'→3', on the bottom strand of the 'phage, short strand of the insert [on which the trinucleotide is d(TTC)], were:

*Xho*I end
ligation-piece sequence

47,1 GAGCTGGCCTCGAACTCGTTCCGA_CTCGACAGACTGAT . . .

47,3 GAGCTGGCCTCGAACTCGTTCCGAGCTCGACAGACTGAT . . .

The sequencing of the right arm of the palindrome of 47,1 was repeated after performing a new amplification of the ligated template DNA in case the first result had been due to a PCR artefact, but the result was the same. All the other 'phage had the correct sequences.

Usually, the results from both stacks are combined for each 'phage without scaling, and when this is done with the d[(GAA)·(TTC)] data (leaving out the aberrant isolate), the result is shown in Figure 7.8a. However, if the results from

stack 2 are scaled up using the reference 'phage results, the picture is as shown in Figure 7.8b, with much smaller 95% confidence intervals.

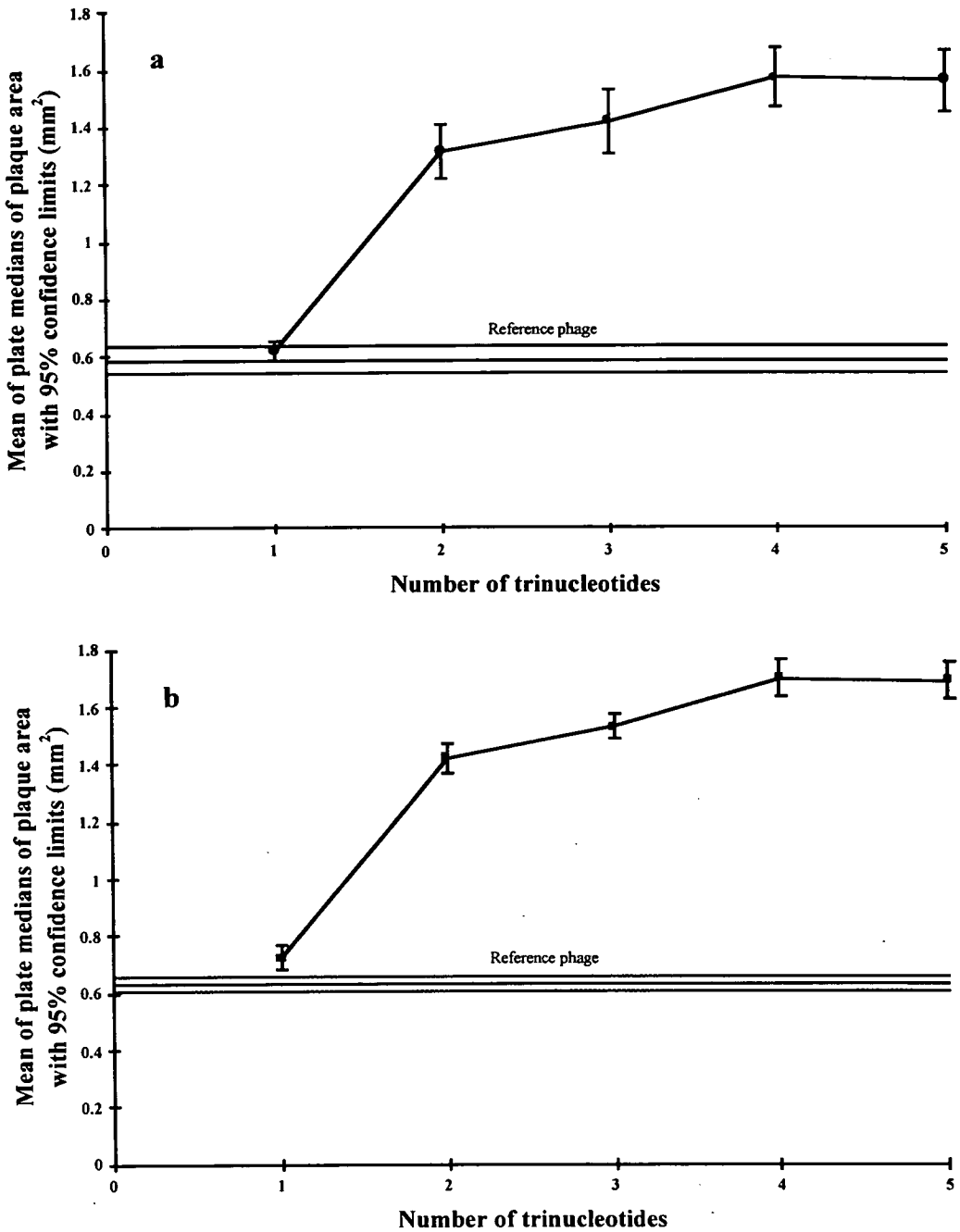


Figure 7.8 Combined plaque size results for all d[(GAA)·(TTC)] 'phage with correct sequences (a) without scaling and (b) with results from plates of stack 2 scaled up to those of Stack 1 using the reference 'phage results.

The plaque size of the 'phage containing one d[(GAA)·(TTC)] trinucleotide in the centre of the palindrome is a little larger than that of the reference 'phage whereas the 'phage containing one d[(CAG)·(CTG)] trinucleotide with the DRL167 construct produced plaques markedly smaller than those of the reference 'phage (see Figure 4.2 or the scaled results in Figure 7.5). However, the inserts used were asymmetrical and the plaque size of 'phage with one d[(CAG)·(CTG)] trinucleotide with asymmetrical inserts of either orientation was more than twice that of the reference 'phage and the plaque size of 'phage containing $d[(CAG)\cdot(CTG)]_2$ was a little larger than that of the reference 'phage (see Figure 7.5) so it is likely that in a symmetrical insert - in the original parent 'phage, DRL167, at least - one d[(GAA)·(TTC)] trinucleotide would produce very small plaques, as expected from the *in vitro* results quoted earlier.

The plaque size for 'phage with inserts of $d[(GAA)\cdot(TTC)]_{1\text{-}5}$ does not zigzag but it does not increase as a straight line either. It shows a new shape, not encountered before. It seems to curve over to a plateau. This is discussed below.

Discussion

The construction of a bacteriophage with a palindrome with an asymmetric centre has enabled it to be determined that the orientation of the insert in the palindrome does not affect plaque size and that the orientation of the centre is not reversing at a rate which would disrupt such an investigation. However, the hope of producing a method by which sequences could be tested for folding potential in the same way as before, but with the facility to sequence them, has only been partially fulfilled. The asymmetry allowing the cleavage on only one side of the insert, and hence the sequencing, has the cost of increasing the palindrome size.

With d[(CAG)·(CTG)]- and d[(GAC)·(GTC)]-containing 'phage the plaque size was seen to increase with increasing numbers of repeat units, though in different ways. No slackening of this increase was seen within the 5 repeat unit length tested

but clearly one could not expect this increase to go on undiminished. The plaques could not increase in size to beyond that attained by 'phage not encumbered with a palindrome at all and therefore one would expect increasing plaque size with a series of 'phage with lengthening inserts of non-palindromic sequence into the palindrome centre to start curving over at some stage towards this plateau size.

In one of the first plaque assays a 'phage without a palindrome, DRL152 (*spi6*, *cI*857, χ^+C153, see Chapter 2) was included, but the plaques were so large that, at the density of plating used, most of the plaques overlapped and could not be measured. Those plaques that could be measured must have been unrepresentatively small because some of the 'phage with the 462 bp palindrome and the longer trinucleotide-repeat inserts made bigger plaques, so, afterwards, DRL152 was assayed again with DRL176 and DRL224 (containing $d[(CAG)\cdot(CTG)]_5$). The old protocol was still being used then and five plates were poured of each 'phage and the median of all plaques of each 'phage was found. The results were DRL176 0·66 mm^2, DRL224 2·27 mm^2, and DRL152 3·52 mm^2. This gives the plaque-area ratios for DRL152/DRL224 as 1·55 and for DRL152/DRL176 as 5·36, where the ratio of DRL224/DRL176 was 3·46.

There was one other occasion when these three 'phage were assayed together. In Chapter 3 it was mentioned that on day 3 of the test of the ratio of plaque sizes of DRL176 and DRL224 under varying conditions, five plates of each 'phage were used and five plates of another 'phage. That 'phage was DRL152. The median plaque areas then were 0·51 mm^2 for DRL176, 1·60 mm^2 for DRL224, and 2.67 mm^2 for DRL152, giving the ratios 1·67 for DRL152/DRL224 and 5·27 for DRL152/DRL176, with the ratio of DRL224/176 3·15. (The means of medians of plaque areas, later calculated from the same data, were very similar: 0·51 mm^2 for DRL176, 1·61 for DRL224, and 2·68 mm^2 for DRL152, giving ratios 1·66 for DRL152/DRL224 and 5·20 for DRL152/DRL176, with the ratio of DRL224/176 3·13.) As results were to show, it was perhaps not the best day to have made the comparison because the ratio between the median plaque sizes of DRL176 and DRL224 was rather higher that day

than on other days (see Table 3.1 and Figure 3.1), the ratio usually being about 2·5 - 2·6.

With the new plaque assay protocol and more plates more accurate results might be obtained but anyway, the results suggest that at around 7 or 8 trinucleotides in the palindrome centre the line would have to start curving over towards the maximum plaque area represented by that of the palindrome-less 'phage. The problem, therefore, with the new parent 'phage with the asymmetric inserts, is that if the plaque areas of all palindrome-containing 'phage are increased because of the asymmetry then the maximum plaque area will be reached with a smaller number of central trinucleotides and there will therefore be less opportunity to observe patterns produced by varying numbers of repeats. There is just a hint that the d[(GAA)·(TTC)]$_n$ data shown in Figure 7.8 might have shown a zigzag pattern, not like that of the d[(CAG)·(CTG)]$_n$ repeats but perhaps a bit like that of the d[(GGC)·(GCC)]$_n$ repeats (see Figure 6.3), had the inserts been made into the original parent 'phage, DRL167. In other words, the d[(GAA)·(TTC)]$_n$ data might have shown a tendency to smaller plaques with odd numbers of repeats than with even numbers but this may have been frustrated by squeezing the whole plot up towards the maximum plaque size by the inclusion of the 1 bp asymmetry.

One would have to check the d[(GAA)·(TTC)]$_n$ repeats in DRL167 or at least with symmetrical inserts in one of the new parent 'phage, and it would be desirable to compare the plaque area of DRL152 with those of some of the asymmetric 'phage. The suggestion that there might be a zigzag pattern with d[(GAA)·(TTC)]$_n$ repeats is not, however, made to suggest that these repeats might form stable hairpins, but the effect of a single d(GAA) trinucleotide making a very stable loop might last for the addition of a few more trinucleotides before it was lost by the instability of a longer and longer length of trinucleotides not much inclined to self-associate.

As noted earlier, DRL167 seems to produce smaller plaques than the new parent 'phage with symmetrical inserts with the same centres and it might be that patterns could be discerned better in 'phage with asymmetrical inserts by using

DRL167. This could be done by making a single base-change in the inserts that would restore one of the *Sac*I sites, instead of knocking out both. This would destroy the *Bsa*I site on the side that the *Sac*I site was restored (unless the inserts were made 1 bp longer at each end), which would not matter, and the 1 bp asymmetry would be separated from the central trinucleotides by 6 bp of symmetrical sequence (see Chapter 4) instead of only 4 bp with the arrangement used in this chapter, so the asymmetry might have less effect in raising plaque sizes. Of course, a new ligation-piece would have to be made for PCR, with a *Sac*I end, but it could have the same primer sequence. The main problem might be that one would have to leave out the recleavage stage after ligating the insert into the palindrome because there would still be a *Sac*I site present in 'phage with inserts. This could mean that 'phage without inserts might outnumber ones with inserts. If this were so, it would mean a lot of extra work purifying isolates and making minipreps of their DNA to find ones with inserts. As mentioned in Chapter 2, when the very first ligations were caried out, with inserts containing $d[(CAG)\cdot(CTG)]_1$ and $d[(GAC)\cdot(GTC)]_1$ into DRL167, a batch of packaged 'phage was made with $d[(CAG)\cdot(CTG)]_1$ without recleaving with *Sac*I. However, after plating, only four plaques were picked from this construction for 'phage selection (4,5-8) and of these there was just one that did not grow on R594 (*rec*$^+$, *sbcBC*$^+$) but did grow on JC9387 (*recBC*, *sbcBC*), *i.e.* had a long palindrome. It was 4,8 and it did prove to have an insert. The tube of packaged 'phage from which it came still exists so more could be plated to see how much trouble not recleaving would cause.

DRL167, however, could not be used for the other purpose for which the new 'phage were designed, namely the screening of random DNA sequences for strong hairpin-forming tendencies. As mentioned in the introduction to this chapter, this requires opposite overhangs when the palindrome centre is cleaved so that a single strand with a degenerate central sequence could be ligated to both sides and then the other strand filled in by polymerase. However, we considered that the largest

practical length of sequence that one could test in this way would be about 6 bp. This would give a little over 2,000 different central sequences.

If one plated the packaged 'phage at a density of about 100 plaques/plate for reasonable separation, one would then need about 20 plates in order for each possible sequence to be represented by an average of one plaque. Many would by chance not be represented. Even if one plated ten times as many p.f.u. (on 200 plates), some 'phage would not be represented. The agar would of course be PSQ agar, and the host strain N2364, so that the plaques of 'phage with good hairpin-forming inserts would make small plaques. Then one would have to pick all the small plaques from all the 200 plates (and there would be several on each just because of some 'phage adsorbing late) and make suspensions from them, and then replate every suspension on a separate plate to check whether it really produced small plaques or whether the small plaque was only the result of late adsorption to the host surface. After incubation, it would be very easy to pick out plates on which nearly all plaques were small from ones on which most were large with just a few small ones, but then from every 'phage suspension identified one would need to prepare a plate lysate from which to make a DNA miniprep. Then with each one would need to go through the process of cleavage, ligation to the end-piece, PCR and DNA purification before running sequencing reactions. All that would only find the sequences forming good hairpin-loops with even numbers of bases in the loop. It would all have to be repeated with seven degenerate bases for the odd-membered loops.

The purpose of making this description is to show why it would only be practical to screen a short length of unknown sequence, and having made that clear, it can be seen that the new 'phage would be suitable for this task because it would still be able to produce plaques markedly smaller than those produced by 'phage with sequences not good at forming hairpins. As seen in this chapter, 'phage containing $d[(CAG)\cdot(CTG)]_2$ or $d[(GAA)\cdot(TTC)]_1$ in asymmetrical inserts produce plaques just a little larger than those of the reference 'phage.

Chapter 8

Concluding Remarks

Summary of conclusions directly from this work

In this work, a plaque area assay, previously developed in this laboratory, has been used to examine the hairpin-forming tendencies of some trinucleotide repeat DNA sequences. A λ bacteriophage construct containing a long DNA palindrome can be propagated on an *sbcC* mutant *E. coli* host but produces much smaller plaques than 'phage not containing a palindrome. (A palindrome of 462 bp was used, along with more palindromic sequence in inserts.) Sequences inserted into the centre of the palindrome will increase the size of plaques formed if they do not tend to form stable hairpin loops at the central axis of the palindrome but will preserve small plaque size, or even diminish plaque size, if they do form stable central hairpin loops.

I have examined the variables in the method to see how they affect the results and revised the protocol to improve accuracy and precision. One of the variables, affecting plaque area, that I could not keep constant, is the degree of drying of the agar which in turn affects the concentrations of salt and nutrients. Plaques are smaller on drier plates. I have shown that, though median plaque area can vary considerably depending upon the condition of the agar, the ratio between the median areas of plaques formed by different strains of 'phage is fairly constant on plates left drying at room temperature for four days or more. This knowledge has been used to scale the results from different assays by the use of a reference 'phage plated in each assay. The principal results obtained with the assay are:

1. The single strands of d(CAG)·d(CTG) repeats have a tendency to form quasi-hairpins *in vivo* and even numbers of repeat units make much more stable structures than odd numbers. This has subsequently been found to be the case *in vitro*.

2. $d(CTG)_2$ and/or $d(CAG)_2$ makes a very stable hairpin loop which is as stable as that formed by d(CTTG) which has been shown to have only two unpaired bases *in vitro*.
3. The single strands of d(GAC)·d(GTC) repeats do not show any marked preference for quasi-hairpin formation with odd or even numbers of repeat units, again agreeing with *in vitro* findings that $d(GAC)_n$ hairpins are about equally likely to form quasi-hairpins with odd or even numbers of repeat units.
4. The loops formed by $d(GAC)_2$ and $d(GTC)_2$ in the centre of a palindromic sequence *in vivo* are less stable than that of the sequence d(AGTTCT), believed to have four unpaired bases *in vitro* and *in vivo*, so may have six unpaired bases.
5. When the single strands of d(CGG)·d(CCG) repeats are constrained to fold in the frame d[CGG)·d(CGG) and d(CCG)·d(CCG) *in vivo* they show a preference for forming quasi-hairpins with even numbers of repeat units, but when constrained to fold in the frame d(GGC)·d(GGC) and d(GCC)·d(GCC) they show a preference for odd numbers of repeat units. Neither of these tendencies is as great as the preference of d(CAG)·d(CTG) repeats for even-membered hairpins and this may well be because the two strands prefer to self-associate in different frames. *In vitro* work by others has shown that the G-rich strand prefers to align in the frame d(GGC)·d(GGC). Alignment of the other strand is still in dispute but most support is for the frame d(CCG)·d(CCG), at least with short tracts.
6. When the strands are constrained to fold in the frame d(GCG)·d(GCG) and d(CGC)·d(CGC) no folding preference is detected with the plaque area assay up to five repeat units.
7. Any possible flipping of the orientation of the central sequence of a long palindrome in a modified λ 'phage grown on *E. coli* strains JC9387 or N2364 by recombination between the palindrome arms is not frequent enough to upset determination of whether orientation of an insert affects results.
8. In the original 'phage construct used, the orientation of inserts could not be determined before or after insertion. I have made new constructs in which the

orientation can be predetermined. DNA recovered from insert-containing 'phage after culture has been sequenced and shown to contain the insert in the orientation in which it was inserted and, by use of the same sequences inserted in either orientation, orientation has been shown not to affect plaque area results used for the above work.

9. Because secondary structure in the long DNA palindrome prevents sequencing, it is necessary to introduce a single base-pair of asymmetry between the two arms of the palindrome so that the DNA can be cleaved on one side of the insert but not the other. This asymmetry does not eradicate the ability of the 'phage construct to distinguish between sequences forming strong and weak central hairpin-loops but it does increase the size of plaques for all central sequences and so probably diminishes the number of repeat units that can be inserted before plaque size becomes so close to that of 'phage containing no palindrome that patterns of variation due to repeat number cannot be detected.
10. The sequence d(GAA) in the palindrome centre appears to form a tight loop as it has been shown to do *in vitro*.

Further work

In this section I list investigations that I should have liked to do to consolidate and extend the work reported. This is mainly to show that I have thought of them and not to suggest that they may actually be carried out.

1. Repeat the d(CAG)·d(CTG) and d(GAC)·d(GTC) investigations using the new protocol, particularly assaying the $d[(GAC)\cdot(GTC)]_{1\text{-}5}$ sequences in the same assay to see whether the plaque area *v.* repeat unit number plot is as straight as it appeared to be, remembering that the plaques of 'phage bearing $d[(GAC)\cdot(GTC)]_1$, $d[(GAC)\cdot(GTC)]_{2\ \&\ 3}$, and $d[(GAC)\cdot(GTC)]_{4\ \&\ 5}$ were measured in three separate assays.

2. Extend both series to at least seven and preferably a few more repeat units to observe what happens to the respective plot shapes in relation to approach of the maximum plaque size produced by a 'phage without a palindrome.
3. Try inserting a non-repetitive non-palindromic sequence into the palindrome centre in 3-bp steps for comparison. We argued with our reviewers that something of this kind would not be particularly helpful to interpretation of the other results. Each series acted as its own control, comparing odd and even numbers of the same repeat and any other sequence would make some sort of secondary structure. It might then be necessary to try more than one non-repetitive sequence, but I should still like to see the effect of steadily moving apart the inverted repeats with unremarkable sequence.
4. Repeat investigation 2 with asymmetric inserts in the new 'phage construct to see how much the latter may reduce the length of repeat tract that can be investigated. Probably less repeats would be needed for this investigation.
5. Try a much longer tract of d(CAG)·d(CTG) repeats, perhaps 20 or 30 units, in the palindrome centre to see whether plaque size would really be large - in agreement with the idea that multiple competing off-centre folding positions in the tract would increase plaque size - or whether plaque size would be small.
6. Try inserting a long, perhaps $d[(CAG)\cdot(CTG)]_{50}$ repeat tract in a 'phage without a palindrome and see whether its plaque size was reduced relative to a 'phage with no such insert.
7. Produce some more isolates of the symmetrical $d[(CAG)\cdot(CTG)]_2$ orientation B construct and assay this 'phage along with those with similar inserts with one and three repeat units to determine whether the shape of the 1-2-3 plot is the same as that in the asymmetrical version of the new construct or the same as that in the (symmetrical) old construct, the difference being that in the asymmetrical new construct the $d[(CAG)\cdot(CTG)]_1$ plaque size is much larger than that of the $d[(CAG)\cdot(CTG)]_2$ whereas in the old construct it was only a little larger.

How may repeat DNA strands and structures move?

1. The problems

a) Moving strands

In Chapter 1, the repeat expansion model of Richards & Sutherland (1994) was mentioned. The authors proposed that if a repeat tract was shorter than an Okazaki fragment the 5' end of the fragment would be 'anchored' by unique sequence and only simple slippage would occur (Figure 8.1a) but that as the repeat tract grew in length there would be an increasing chance that two single-strand breaks (the ends of an Okazaki fragment) would occur within the repeat tract. They imagined that if a fragment was composed exclusively of trinucleotide repeats it might be able to 'slide' on its template (Figure 8.1b) and that subsequent repair would lead to expansion.

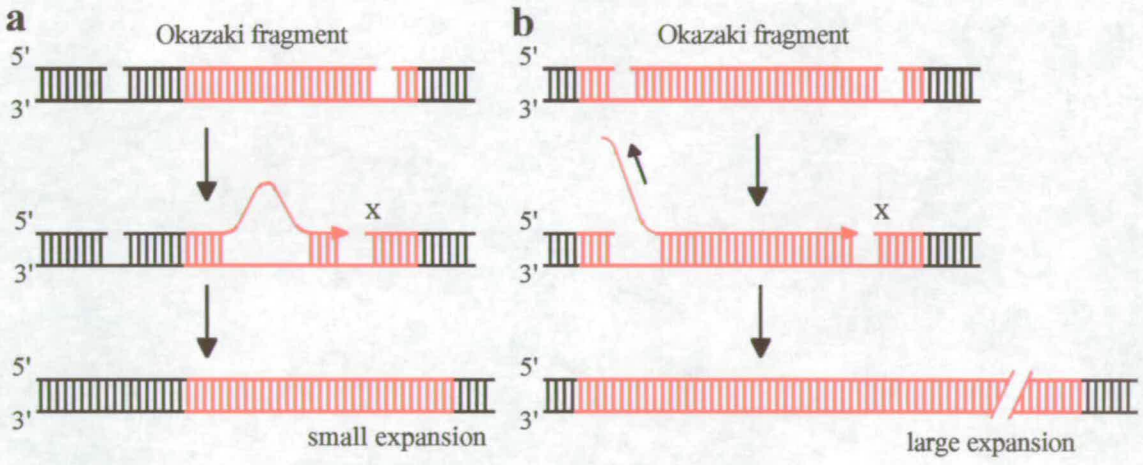


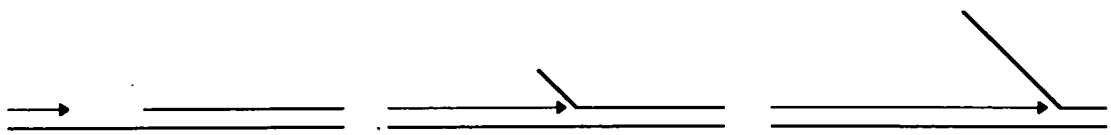
Figure 8.1 The repeat tract expansion models proposed by Richards & Sutherland (1994) where (a) only one single-strand break occurs within the repeats and (b) both ends of an Okazaki fragment lie within the repeat tract. Redrawn. Unique sequences are here shown in black; repeats in red.

There were several deficiencies in this model. It was not explained why the 5' end of an Okazaki fragment would melt off its template (b) if in a repeat tract yet would not do so if it contained any unique sequence (a), nor how repair would lead to

the large expansion proposed. The authors also appeared to believe that the 5' end of a strand could be extended (x marks the spot). A more important problem was that this model would apply equally well to any short tandem repeat and did not explain why certain repeats were much more unstable than others. However, what we (Darlow & Leach, 1995) objected to most was the idea that an Okazaki fragment could 'slide' on its template. The fragment, after all, is not only wound round its template in a double helix but bound to it by a large number of hydrogen bonds, all of which would have to be broken, and what would be the driving force propelling the fragment in a 5' direction? We suggested that such apparent sliding was more likely to occur by the mechanism illustrated in Figure 1.2a (p. 38). In this model, initially cruciform formation by the nascent strand and its template occurs within the repeat tract and then, when the template strand has the less stable of the two hairpins, the hairpin on that strand melts, leaving an intact hairpin on the nascent strand, and the 3' end of the nascent strand is again extended over the same stretch of template as before. This is similar to (a) of Richards & Sutherland (1994) but with the looped-out DNA stabilised by internal bonding, thereby explaining to some extent the movement of the strand and explaining why certain repeats are much more prone to expansion than others. It could of course happen just as easily whether the 5' end of the Okazaki fragment had any unique sequence or not and we proposed that a different explanation of a possible threshold tract length above which expansions become much larger might be related to the minimal length of homology required to initiate homologous recombination, as discussed in Chapter 1.

In their next review including discussion of mechanisms of repeat expansion, Richards & Sutherland (1997) concentrated on gene conversion and mentioned slippage but not their 'sliding' model. However, in the meantime Gordenin *et al.* (1997) proposed another mechanism (which, as mentioned in Chapter 1, had striking similarity to something proposed by Ripley (1982), Figure 1.1, p. 34). As Gordenin *et al.* (1997) relate, when an Okazaki fragment is extended up to the downstream

fragment the polymerase pushes the 5' end of that fragment off the template, creating a 'flap' like the free end in Figure 8.1b but without any sliding of that fragment:



Normally this flap is chewed off by a protein known as RAD27 in yeast and FEN1 in mammals. In mutants, large duplications are found at about 1,000-fold the normal rate. There is also an increased rate of recombination and mutants cannot survive if unable to repair double-strand breaks. FEN1 can only act on single-stranded DNA and a partially double-stranded flap blocks its action. The T5 'phage homologue has been shown to be like a bead with a fine bore which will just allow a single strand of DNA to be threaded through it but not a double strand. Gordenin *et al.* (1997) suggested that a flap composed of one of the disease-causing trinucleotide repeats could fold over to pair with itself, forming a hairpin, or with the double-stranded DNA, forming a triplex. Then the 5' end of the hairpin might be ligated to the upstream Okazaki fragment, leading to expansion after another round of replication, or the template might be cleaved and the gap in it filled in opposite the ligated flap DNA, or the secondary structure might be cleaved, leaving a double-strand break, and repaired by recombination with the double-stranded homologue.

Following this, Sutherland *et al.* (1998) have now resurrected their sliding Okazaki fragment model. They have corrected the error of the 5' extension of a strand (at x in Figure 8.1 a & b) and have incorporated the folding of the flap to protect it from FEN1, but they still talk of the flap being created by the Okazaki fragment sliding in a 5' direction along its template with no suggestion as to how or why this might happen. I shall return to this after presenting some other apparently unworkable suggestions.

b) Moving hairpins

In Chapter 5 it was mentioned that Chen *et al.* (1998) had put forward an idea that would be an explanation of the observation that we made (Darlow & Leach, 1998a) that the data of Smith *et al.* (1994) showed that $d(CCG)_{11}$ constrained to pair in frame 1 was not quite as good a substrate for methylation as unconstrained $d(CCG)_{15}$. It has already been mentioned that the human DNA-(cytosine-5)-methyl-transferase recognises the motif $\begin{matrix}\text{-CpG-}\\ \cdot \\ \text{-C-}\end{matrix}$. Smith *et al.* (1987) found that the enzyme methylated the cytosine of the CpG dinucleotide seven times more rapidly if it was C·C mispaired than if it was C·G Watson-Crick bonded and Smith (1991) suggested that, driven by the fact that unusual DNA structures can cause mutation, the enzyme might have been evolved to recognize unusual DNA structures and methylate them so that they would be recognized by proteins that bind methylated DNA and return the structures to normal duplex form. Smith *et al.* (1994) suggested that methylation of d(CGG)·d(CCG) repeat tracts, and other sequences capable of forming hairpins containing mismatched cytosines, might be achieved *via* melting of the duplex, formation of a hairpin by the C-rich strand, methylation of the mismatched CpG cytosines in the hairpin, return of the C-rich strand to duplex with its complementary strand, and then rapid methylation of that complementary strand, being now part of a hemimethylated duplex.

Chen *et al.* (1998) elaborated upon this idea. It was mentioned in Chapter 5 that they performed *in vitro* DNA synthesis on single-stranded $d[C(GGC)_n]$ and $d[G(CCG)_n]$ templates in M13 in a PCR machine with extension temperatures ranging from 45 - 85°C (the annealing temperature was not given). Though they did not try extension at 37°C, they concluded from their results that $d(GCC)_n$ will not form a hairpin in the presence of its complementary strand unless n = some number greater than 8 but less than 21 and that $d(GGC)_n$ will not form a hairpin in the presence of its complementary strand even for n = 21. They developed the hypothesis that during replication of long $d[(CGG)\cdot(CCG)]_n$ tracts three-way junctions will be formed consisting of hairpins of the C-rich strand sticking out from

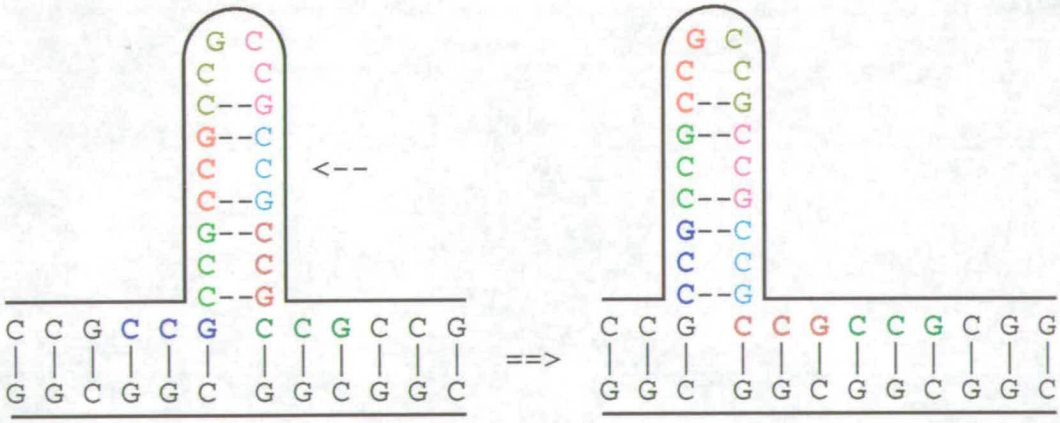
the complementary duplex. The hairpins, they maintain, would have the facility to change in two ways, increasing in length, which they call 'slipping', and moving sideways along the complementary duplex, which they call 'sliding'. This they suggest is the basis of the hypermethylation of expanded d(CGG)·d(CCG) repeat tracts: the mismatched cytosines on the hairpin of the C-rich strand (paired always in frame 1) would be methylated and then the methylated cytosines would be laid down into the duplex as the hairpin moves sideways, thus producing a hemimethylated duplex which would rapidly be fully methylated.

Chen *et al.* (1998) backed up this hypothesis with methylation results from three different classes of substrates: 'completely mobile' junctions, made from annealing of $d(GGC)_n$ and $d(GCC)_{n+m}$ oligonucleotides, 'partially mobile' junctions, in which the ends of the complementary strands were joined by T_4 loops, and an 'immobile' junction, in which the C-rich strand hairpin is also closed by a T_4 loop. (The 'partially mobile' and 'immobile' junctions were formed from single oligonucleotides with the ends meeting at a point on the complementary duplex part of the structure that effectively had a nick there.) The authors found that the 'partially mobile' structure with the same number of bases as the one 'immobile' structure had an approximately nineteen-fold greater rate of methylation. The smallest 'completely mobile' structure had 50% more bases and its methylation rate was 25·4 times that of the 'immobile' structure. (With both the partially mobile and the completely mobile structures the relative rates increased with the length of the C-rich hairpin.) The rate of methylation of $d(GGC)_{10}·d(GCC)_{21}$ was more than double the rate for hairpins of $d(GCC)_{21}$ alone, showing that the rate for the former was not due to complete dissociation and methylation of the $d(GCC)_{21}$ hairpin (hairpins of $d(GGC)_n$ are methylated very little).

The largest of these structures (a completely mobile one) was made from $d(GGC)_{15}$ and $d(GCC)_{21}$ which would give, when as near symmetrically annealed as possible, two complementary arms of 22 bp and 23 bp and one d(GCC) hairpin containing 6 repeats and perhaps it could change shape by at least partial melting and

reannealing. The concept of 'sliding' of a hairpin along a complementary duplex has two major problems. One of them was considered by Chen *et al.* (1998). It was that the hairpin must rotate relative to the axis of the complementary duplex as it moves along. Indeed, it must rotate 360° for every $10^1/_2$ bp (= $3^1/_2$ repeats). Chen *et al.* (1998) did not say this directly but remarked that "a migrating three-way junction is likely to cause negative supercoils behind RNA polymerase" and, of course, with the hairpin not rotating there would also be positive supercoils produced in front. Chen *et al.* (1998) only said that perhaps the negative supercoiling "is accommodated during transcription by the normal length of the repeat". They did not speculate about what energy source would propel a hairpin along the duplex when methylation was occurring.

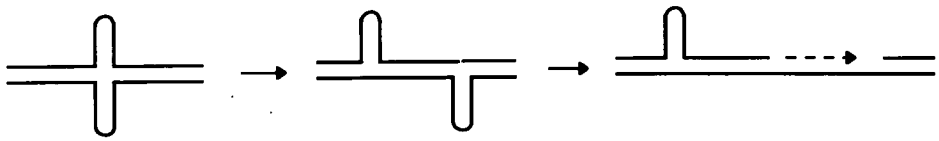
The other problem was not mentioned. It is that a hairpin cannot migrate along a duplex, even by one repeat, without every hydrogen bond in the whole hairpin being broken and new bonds being made with bases 6 nt further along the opposite side (yes, 6, not 3; look at the light green and the magenta repeats):



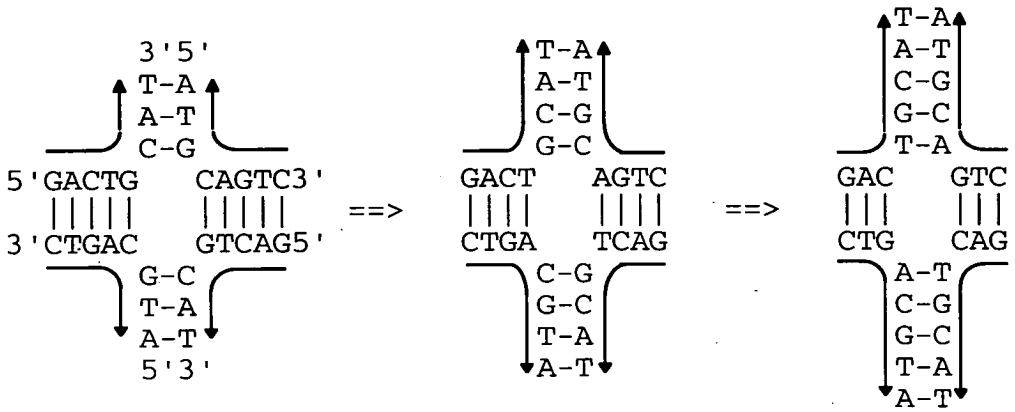
In reality movement would be even more difficult than it appears from this diagram because the junction would actually be a Y shape, with all three angles 120°, and all three arms would have a double helical twist and the hairpin would not be able to unwind in order to allow its two sides to pass each other in opposite directions.

Petruska *et al.* (1998) have fallen into the same trap. As mentioned in Chapter 1, they proposed that expansion might occur by the following process. First

a cruciform might be formed in the repeat tract consisting effectively of a hairpin on each of the two complementary strands, the hairpins opposite one-another. Then the hairpins migrate apart, moving in opposite directions along the complementary duplex. Then one of the strands is nicked opposite to a hairpin on the other strand, the hairpin is unfolded, opening up a gap in the nicked strand, and this gap is filled in by polymerase and ligated, thereby achieving expansion.

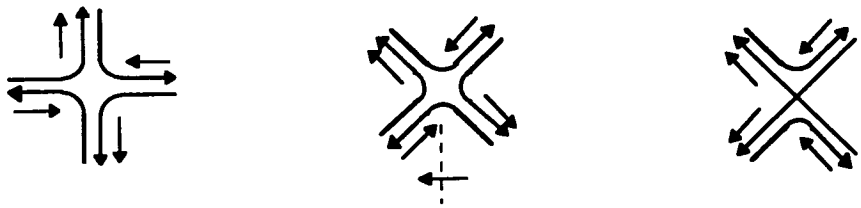


The problem illustrated of movement of three-way junctions does not arise with four-way junctions. Consider first cruciform extrusion, illustrated below.

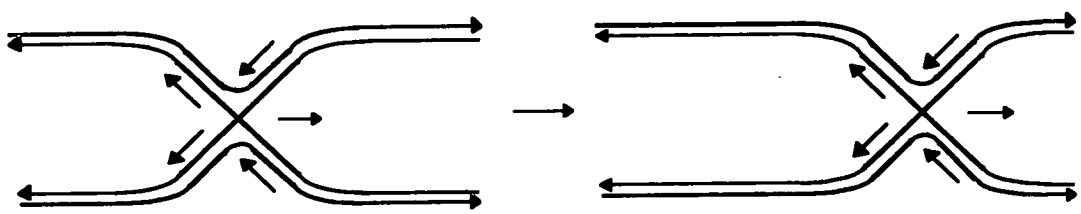


Two double strands go into the junction and two double strands come out. The arrows here indicate the direction of movement, not 5'→3'. This is only a diagram of course, and really each of the four arms of the junction has a double helical twist, but no strands have to slip past each other as in the hypothetical three-way junction movement.

In this junction, as drawn, the two strands moving into the junction are opposite one-another and the two strands moving out of the junction are opposite one-another. If we twist two adjacent arms of the junction about a diagonal axis between them, we have the familiar Holliday junction.

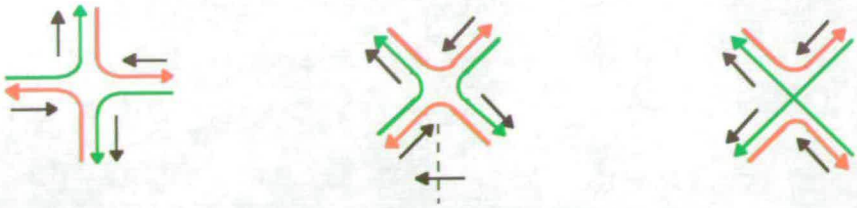


Here the arrows on the strands indicate 5'→3' and the black arrows indicate the movement of the double-stranded DNA arm. Strands the same colour have the same sequence and strands of different colours are complementary. In the diagram on the right the two arms moving into the junction are adjacent and the two moving out are adjacent. Now we can either think of the junction remaining static and the strands moving through it, or the strands remaining static and the junction moving along them.

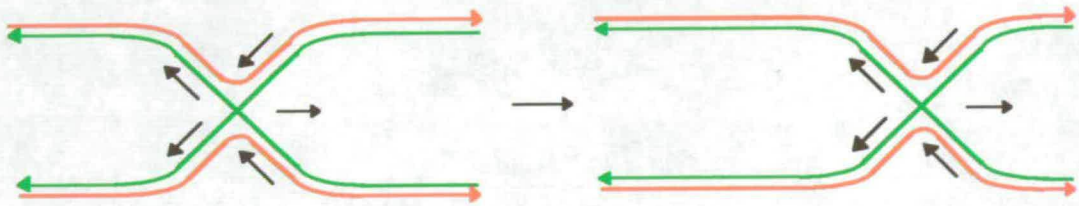


This is branch migration. What happens at the junction is identical to cruciform extrusion. The only overall difference is that in the latter case two of the arms end in hairpin loops. The four arms of the junction remain opposite one-another. The four-way junction does not break up into two three-way junctions as in the scheme of Petruska *et al.* (1998), which those authors called branch migration.

Just as the suggestion of Chen *et al.* (1998) that a hairpin of trinucleotide repeats might slide along a complementary duplex could not be realized without breaking all the bonds in the hairpin, so their other suggestion, that the hairpin could elongate, could not be achieved without breaking all the hydrogen bonds in the complementary arms of the three-way junction to move more repeats into the hairpin. Sinden & Wells (1992) suggested a mechanism of trinucleotide repeat expansion involving hairpin growth. They envisaged that if the advancing polymerase met a strong block (which might be protein tightly bound to repeat tract,



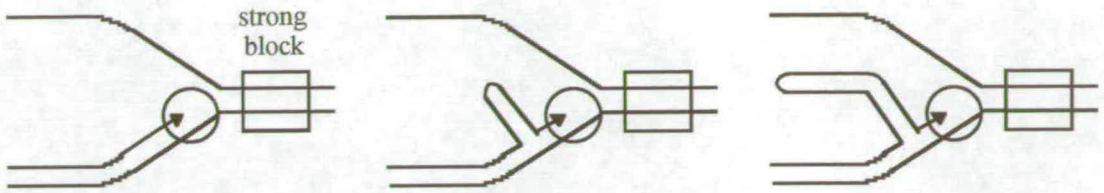
Here the arrows on the strands indicate 5'→3' and the black arrows indicate the movement of the double-stranded DNA arm. Strands the same colour have the same sequence and strands of different colours are complementary. In the diagram on the right the two arms moving into the junction are adjacent and the two moving out are adjacent. Now we can either think of the junction remaining static and the strands moving through it, or the strands remaining static and the junction moving along them.



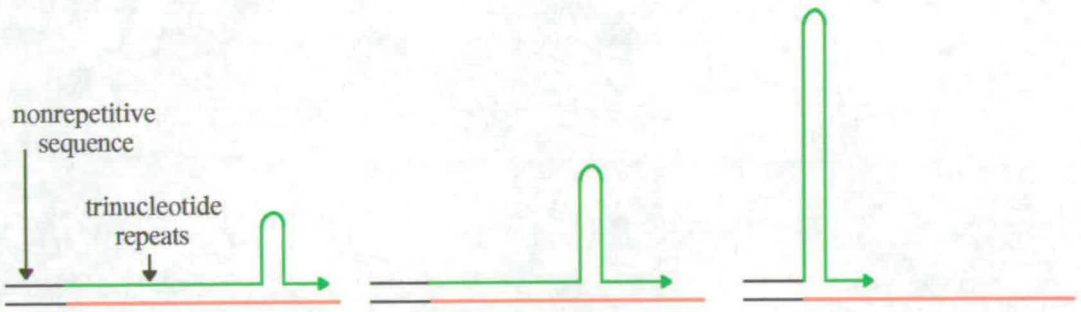
This is branch migration. What happens at the junction is identical to cruciform extrusion. The only overall difference is that in the latter case two of the arms end in hairpin loops. The four arms of the junction remain opposite one-another. The four-way junction does not break up into two three-way junctions as in the scheme of Petruska *et al.* (1998), which those authors called branch migration.

Just as the suggestion of Chen *et al.* (1998) that a hairpin of trinucleotide repeats might slide along a complementary duplex could not be realized without breaking all the bonds in the hairpin, so their other suggestion, that the hairpin could elongate, could not be achieved without breaking all the hydrogen bonds in the complementary arms of the three-way junction to move more repeats into the hairpin. Sinden & Wells (1992) suggested a mechanism of trinucleotide repeat expansion involving hairpin growth. They envisaged that if the advancing polymerase met a strong block (which might be protein tightly bound to repeat tract,

or might be a quadruplex or a triplex) then it might repeatedly slip backwards, building up an ever-lengthening hairpin of repeats behind it on the nascent strand:



It has already been mentioned that it was pointed out that mutation usually occurs on the lagging strand, not the leading strand. The point I want to make here is that, unless every bond in the hairpin is broken and remade with a new base, a hairpin on a single strand cannot grow unless it recruits more single strand from both sides. Therefore in the situation above, redrawn from Sinden & Wells (1992), the hairpin would not stay in the same place but move in a 5' direction as it grew -

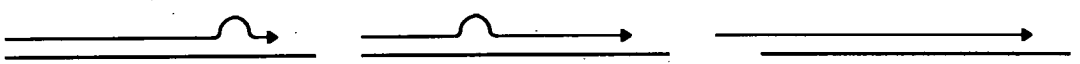


- and once its base was at the 5' end of the repeat tract it would not be able to grow anymore, though a second hairpin might build up against the first if the polymerase slipped back again after further extension.

We have seen that it has been proposed that one strand of a DNA duplex may slide along the other (Richards & Sutherland, 1994; Sutherland *et al.*, 1998) and that a hairpin formed from a single strand of trinucleotide repeat DNA may slide along a complementary duplex (Chen *et al.*, 1998; Petruska *et al.*, 1998), neither of which appear to be possible. But are they in fact possible?

2. Resolution

In Chapter 1 it was mentioned that it had been found that long strands of tandem repeats could be produced by polymerization starting from short complementary primers. This process was used to produce trinucleotide repeat DNA, and hence RNA for translation, to discover the genetic code (Khorana *et al.*, 1966). In Chapter 4 it was mentioned that Schlötterer & Tautz (1992) reinvestigated the process, synthesizing all ten types of trinucleotide repeat starting from complementary primers of 3 and 5 repeat units respectively and using polymerases active at 37°C and incubating at constant temperature. All continued to grow, only slowing down as reagents ran out, and when the products were used to start new reactions they grew again as rapidly as before. From their investigations the authors concluded that after melting of the 3' end of a strand from its template and reannealing further back, the little bulge looped out behind it could move as a wave in a 3' → 5' direction along the nascent strand to come off at the other end if not tethered:



Schlötterer & Tautz (1992) found that d(GCC)·d(GGC) repeats grew at a barely detectable rate and d(CAG)·d(CTG) repeats very slowly compared with most of the others. From the above discussion, the reason for this now seems immediately obvious. With these repeats, if the little bulge contained more than one repeat unit it would be able to form internal hydrogen bonds and these would severely retard its movement. With repeats not prone to self-annealing the bulge might contain two or three, or possibly even four repeat units, and so the sliding of the strand could be much more rapid. (Probably the looped-out DNA would not be much longer than this otherwise there could be increasing difficulties of the two sides of the loop slipping past one-another even without much tendency for bonding.) Thus we see

that a strand may slide on its template, but the mechanism is least likely to account for expansion of the repeats for which Richards & Sutherland (1994) proposed it.

It could be objected that slippage by any whole number of repeats can occur in any 'frame' where the number of frames (here nothing to do with coding) equals the number of bases in the repeating unit, three for trinucleotide repeats. Thus if, for instance, in the synthesis of a d(CTG) strand, slippage occurred by two or three repeats in the frame (TGC) or (GCT) there would be no tendency for a mini-hairpin to form. However, the little bulge or ripple in the nascent strand would only have to move by one or two nucleotides before its sides would be aligned as (CTG)·(CTG) and become prone to bond. The fact that with the single strands of d(CGG)·d(CCG) repeats bonding might occur, even if not very stably, in at least two of the three frames, might explain why Schlötterer & Tautz (1992) found these strands to slip at the slowest rate of all.

One might invoke the migration of a bulge to permit the sideways movement of a hairpin but it would not be likely to occur spontaneously. The bulge would appear at the base of the hairpin on one side if the hairpin was forced across from the other side. Either the hairpin might be pushed across by 1 - 2 nt and a one-trinucleotide bulge might be formed which would move up and over and down the other side of the hairpin, and disappear again at the base on the other side, allowing the hairpin to move the remaining 1 or 2 nt into position, or the hairpin would have to be forced over by a whole repeat at once, becoming at the same time a whole repeat shorter in height with two repeat units bulged out and this bulge would just move up to the top of the hairpin and there disappear, becoming the new loop. It was mentioned in Chapter 4 that Pearson & Sinden (1996), Pearson *et al.* (1998a) made multiple forms consisting of duplex trinucleotide repeat DNA with hairpins and other structures looped out on either strand at different positions by melting and reannealing. They found that these forms were extremely stable and did not readily change from one into another which indicates that indeed hairpins are not mobile as Chen *et al.* (1998) believe. Of course, in the cell, there are many proteins which

manipulate DNA so one cannot completely rule out such a process but it would surely be very expensive in terms of energy compared, for instance, with movement of a Holliday junction.

Increase in length of a hairpin might be a different matter. A hairpin trapped on a complementary duplex of repeats during replication could become a sink that absorbed small bulges that migrated into it from one-unit slippages by the polymerase as suggested by Harvey (1997). On arrival at the hairpin such a bulge might migrate up to the top and there disappear, becoming the new loop.

Mechanisms of expansion

The literature on trinucleotide repeat expansion disorders is vast and many hypotheses for the mechanism(s) have been proposed. Much work investigating the expansion process has been published since the start of this project that has not so far been discussed. The main areas of enquiry have been: analysis of expansion and deletion products of trinucleotide repeats in wild-type and various mismatch-repair, recombination and other mutants of *E. coli*, including the effects of orientation with respect to the origin of replication and interruptions to the repeat tract; similar work in yeast; observations of somatic and germline expansion and contraction in humans, including the influence of directly flanking sequence, haplotype, and normal allele; and a smaller amount of work so far on instability in transgenic mouse models of trinucleotide repeat disorders as well as work on naturally occurring unstable repeat sequences in mice. There has also been work on mismatch-repair-deficient human cell lines and tumours and entirely theoretical work. It has become quite clear that one cannot accept conclusions at face value; it is necessary to read papers on the contrasting results of different laboratories in detail before deciding whether one can accept the results or the conclusions and there is neither the time nor the space to do justice to this work here. My work has been on DNA structure and I have concentrated on papers on structure but the purpose of establishing whether expanding repeat sequences can form structure *in vivo* is of course to help to

understand the expansion mechanism(s) and I feel I cannot finish this report without making some reference to these investigations. I present some facts and topics of debate have emerged.

1. Structure is important

As discussed in Chapter 1, some short tandem repeat sequences are much more unstable than others and, though not all of the very unstable ones have been investigated for ability to form unusual secondary structure, and though not all of those that have been shown to form unusual secondary structure have been proved to do so *in vivo*, it seems probable that they do, and that this is in some way involved in the expansion mechanism. The unusual secondary structure does not have to be a hairpin.

2. There is more than one mechanism of instability

It seemed clear from the time of the early discoveries of repeat instability in trinucleotide repeat expansion disorders and non-polyposis colon cancer that different factors were involved since instability occurred at a single locus in each trinucleotide repeat disorder and at many loci in colon cancer, but Goellner *et al.* (1997) felt that there was sufficient doubt to warrant investigating this to make sure. Their doubts stemmed from the finding that large increases in repeat tract length could occur in colon cancer (Shibata *et al.*, 1994) and that small changes in repeat number are commonly seen in trinucleotide repeat disorders, both in intergenerational transmission and in somatic cells. They looked at instability in the same 7 trinucleotide repeat loci [4 of d(CAG)·d(CTG) and 3 of d(GGT)·d(ACC) repeats] and 3 dinucleotide repeat loci [all of d(CA)·d(TG) repeats] in Huntington disease (HD) and in non-polyposis colon cancer (familial and sporadic). In both groups they looked at somatic tissue. In HD they compared repeat numbers in DNA from leukocytes in parent-child pairs and in colon cancer they compared DNA from

tumour tissue and normal colonic mucosa from the same individuals. They found that changes in HD were almost entirely confined to the disease locus (known as *IT15*) with the previously-observed bias towards expansion, which was up to 54 repeat units. In contrast, changes in colon cancer occurred at most loci examined, were about evenly balanced between expansion and contraction, and were nearly all in the range –7 to +6 repeats, though there were two expansions of 10 and one each of 14 and 15 units, all at dinucleotide repeat loci. Also there was little evidence of gender bias in colon cancer compared with the previously-observed bias towards expansion in male transmission in HD.

Unfortunately only average starting lengths were given for the repeat tracts at all loci. The *IT15* repeat tract was presumably not nearly as long in the colon cancer patients as in the HD patients so one does not see what would have happened if it had been closer to the length at which instability occurs in HD. The results do however show that the HD patients did not have mismatch repair defects, or at least not any of the ones that occur in colon cancer. Therefore, as had already been deduced, the expansions seen in repeat expansion disorders must be engendered by the specific nature of the sequence. It has been seen that the sequence must not only be one of a particular group of sequences but have beyond a certain length that, for trinucleotide repeats in humans, though not for highly unstable minisatellites, must be free of sequence imperfections.

Kramer *et al.* (1996) investigated the length of the repeats at the *DM* and *FRAXA* loci in mismatch-repair-negative (*hHSH2* and *hMLH1*) and -positive tumour cell lines and in lymphoblast cell lines from individuals with DM full- and pre-mutations, FRAXA full- and pre-mutations, and individuals normal at these loci. They too concluded that mismatch repair deficiency did not cause large changes and that there must be two mechanisms. (They remarked that the results might have been different if only they had had mismatch repair deficient lines with DM and FRAXA repeat tracts in the affected ranges but that to their knowledge none existed.)

The same conclusion was reached by Wells and colleagues (Kang *et al.*, 1996) from observations of interrupted repeat tracts cloned in *E. coli*. A d(CAG)·d(CTG) tract containing 175 trinucleotides, including two d(TAG)·d(CTA) interruptions, 41 trinucleotides apart in the distal end of the tract with respect to the origin of replication, was found to expand producing products with more interruptions and these were also 41 trinucleotides apart. From this the investigators concluded that expansion had occurred as a single large event rather than by an accumulation of small slippages, which would have placed the variant trinucleotides different distances apart, and therefore that the mechanism was different.

3. Normal repair mechanisms may cause instability

The above does not, of course, mean that mismatch repair is not involved in the generation of large expansions in repeat expansion disorders. One could perhaps entertain three possibilities: (i) expansion occurs at a time when DNA repair mechanisms are already very busy and some problem caused by the unstable tracts exceeds their ability to cope and so mutations go uncorrected; (ii) expansion is a consequence of a normal repair mechanism interacting with an abnormal structure formed by the repeat tract; (iii) repair is not involved, *i.e.* the problem caused by the repeat tracts is not amenable to normal repair mechanisms. Of these, (i), which was suggested by Kunkel (1993), is unlikely because in this case one would expect a higher rate of mutations at other loci in people with trinucleotide repeat disorders than in people with normal repair and normal alleles at the trinucleotide repeat loci. Goellner *et al.* (1997) did not make this comparison but the frequency of changes at the other loci in the HD patients was low and the authors did look for an increased incidence of cancer in the HD patients and did not find it.

As to the distinction between (ii) and (iii), Goellner *et al.* (1997) suggested that trinucleotide repeat hairpins are not repaired, but there are indications that normal mismatch repair is required for large changes in repeat tract length. Before

discussing these results from work on *E. coli* and *Saccharomyces cerevisiae* some prefacing remarks seem worth making:

(a) For a number of reasons including the fact that long repeated sequences tend to be reduced in length rapidly when cloned in prokaryotes, work with *E. coli* cannot be expected to provide all the answers to the repeat expansion disorder puzzle, but the organism is easy to use and can still tell us a lot. Also, in the now completely-sequenced genome of *Saccharomyces cerevisiae*, pure trinucleotide repeat tracts of all types longer than 5 units are rare, longer tracts nearly all being interrupted ones (Richard & Dujon, 1996) suggesting that they too may not be tolerant of longer pure tracts.

(b) As was mentioned in Chapter 1, it has been shown that replication slippage between non-adjacent direct repeats brought into proximity by hairpin formation by the intervening DNA tends to occur on the lagging strand. Trinh & Sinden (1991) suggested that this would be so because the lagging strand can be single stranded up to the length of an Okazaki fragment. They were able to demonstrate it in *E. coli* by designing asymmetric sequences, containing palindromes, in which deletion between the repeats would only occur in one orientation with respect to the direction of replication. The principle has since been supported by results of others (Rosche *et al.*, 1995; Pinder *et al.*, 1998). If then large changes in number of trinucleotide repeats are brought about by secondary structure formation and one strand forms a more persistent structure than the other, one would expect the frequency of large deletions to be dependent upon the orientation of the repeat. This might be expected not to apply to expansions because they would be produced by structure formation on the nascent strand and perhaps this might happen just as easily on a leading- as on a lagging-strand template. However, results from *E. coli* by Wells and colleagues (Kang *et al.*, 1995; Ohshima *et al.*, 1996c; Kang *et al.*, 1996; Shimizu *et al.*, 1996) and others (Samadashwily *et al.*, 1997) and from yeast (Maurer *et al.*, 1996; Freudenreich *et al.*, 1997; Miret *et al.*, 1998) show that both events are orientation-dependent in wild-type cells and occur on the lagging-strand arm of the replication fork. With

d(CAG)·d(CTG) repeats deletion occurs when the CTG strand is the lagging-strand template and expansion when it is the leading strand template, *i.e.* the nascent strand on lagging template. With d(CGG)·d(CCG) repeats deletion was found to occur when the CGG strand is the lagging-strand template (Shimizu *et al.*, 1996). This is interesting since, as discussed in Chapter 5, CCG repeats have the greater tendency to form hairpins. It suggests that quadruplex formation by the CGG strand may be the important factor. One trouble with this is that Usdin & Woodford (1995) found that an AGG interruption had no detectable effect upon the stability of $d(CGG)_n$ quadruplexes.

(c) The changes seen in mismatch repair deficiencies result from failure to repair the effects of replication slippage. Since, as we have seen, slippage can occur on any short tandem repeat sequence, whether it has a tendency to secondary structure formation or not, and since expansion and deletion are of about equal frequency, one would not expect orientation of a repeat sequence to affect the frequency of changes brought about by mismatch repair defects.

From results with *E. coli*, Wells and colleagues (Jaworski *et al.*, 1995; Wells *et al.*, 1998) concluded that mutations in the mismatch repair genes *mutS, mutL* and *mutH* stabilised d(CAG)·d(CTG) repeats but Schumacher *et al.* (1998) conversely concluded that intact mismatch repair stabilizes these repeats. This confusion seems to result from the different sizes of repeat change being observed. Schmidt *et al.* (from this laboratory, paper submitted) have reconciled these findings. They have examined changes in the distribution of mutant lengths from a starting tract of $d[(CAG)\cdot(CTG)]_{43}$ at intervals up to about 140 generations. Over time the proportion of plasmids with the starting length diminishes and the proportions with other lengths increase. In wild-type cells a range of 5 - 69 repeats was reached but the rate of dispersion from the original length was greater when the lagging strand template carried CTG than when it carried CAG and this difference was accounted for by the number of plasmids with deletions of >7 repeat units. In mismatch-repair-deficient cells the repeat tract was more unstable and this was independent of

orientation. The increase in instability was mainly due to changes of +1 and −1 repeat from 43 which were barely seen in wild-type as they are most efficiently detected and corrected by the repair system. This however was about the same in either direction and did not therefore account for the loss of orientation-dependence. The latter was accounted for by a greater decrease in deletions of >7. Large changes in either direction were attained by gradual accumulation of small changes. These results therefore indicate that large deletions are orientation- and mismatch-repair-dependent and are eliminated by mismatch repair deficiency. In wild-type cells a slightly greater proportion of expanded plasmids was seen when the lagging-strand template carried the CAG repeats, *i.e.* when CTG was on the nascent strand, than *vice versa*, and this difference was maintained in the mismatch-repair mutants, the increase in small changes in both directions being superimposed upon it.

In yeast, Heale & Petes (1995) reported that stabilization of $d[(GT)\cdot(AC)]_n$ tracts by variant repeats requires a functional mismatch repair system, but Petes *et al.* (1997) retracted this, saying that variant repeats stabilized the tract regardless of mismatch repair status. However, the latter did find that in mutants of *pms1* (a *mutL* homologue) and *msh2* and *msh3* (*mutS* homologues) changes were exclusively of one or two repeat units lost or added whereas in wild-type and *msh6* mutants (which with these tracts behaved as wild type) about 50% of the changes were large (up to 28 repeat units) so these results again indicate that functional mismatch repair (of some kinds at least) is necessary for large changes. (Interestingly, as with the *E. coli* work (Kang *et al.*, 1996), variant repeats were included in the altered regions, all the deletions causing loss of a variant repeat and some of the expansions causing introduction of a copy of the variant.) In contrast, (Schweitzer & Livingston, 1997) using d(CAG)·d(CTG) tracts of up to 92 units, found that mutations in the same three mismatch repair genes (*PMS1*, *MSH2* and *MSH3*) only introduced small changes in repeat number while not diminishing the numbers of large deletions at all. Likewise, Miret *et al.* (1997), with the same type of repeat and mutations in the same genes, found no effect of mismatch repair in an assay which could detect only

deletions of 12 or more repeats (into a range of 8 - 38 repeats from 50) and subsequently (Miret *et al.*, 1998), using an assay that could only detect expansions of 5 or more repeats from a starting tract of 25, found no effect of *MSH2* deficiency.

As to how intact mismatch repair might cause deletion, Jaworski *et al.* (1995) suggested that removal of a section of the nascent strand following a small slippage event, leaving the template single-stranded, would create the opportunity for the latter to form secondary structure. This suggests that secondary structure does not have an opportunity to form on the lagging strand when it is looped out from the replication complex prior to starting a new Okazaki fragment, presumably because it would always be covered by single-strand-binding protein. Pearson *et al.* (1997) made S-DNA (as described in Chapter 4) with d(CAG) and d(CTG) oligonucleotides both of 30 repeat units or both of 50 and also 'SI-DNA' ('slipped intermediate DNA'), in which one strand is longer than the other, and found that human MSH2 bound to these structures and the binding efficiency showed that the protein binds better to CAG repeat hairpins than to CTG hairpins. The authors suggested that this might either mean that CTG hairpins escape detection more often than do CAG hairpins or that MSH2, alone or in a complex, might protect CAG loops from repair, thereby allowing replication to finalize expansion. They did not notice that the evidence suggests that it is CTG hairpins that cause expansion.

Finally, Wöhrle *et al.* (1995) found somatic instability of expanded DM repeats in foetal tissues and cultured cells compared with somatic stability of expanded FRAXA repeats. They concluded that the methylation in the latter allowed efficient strand-specific methyl-directed mismatch repair and that the instability in the former was due to misdirected repair due to the lack of methylation. This, of course, presupposes that mismatch-repair is methyl-directed in humans.

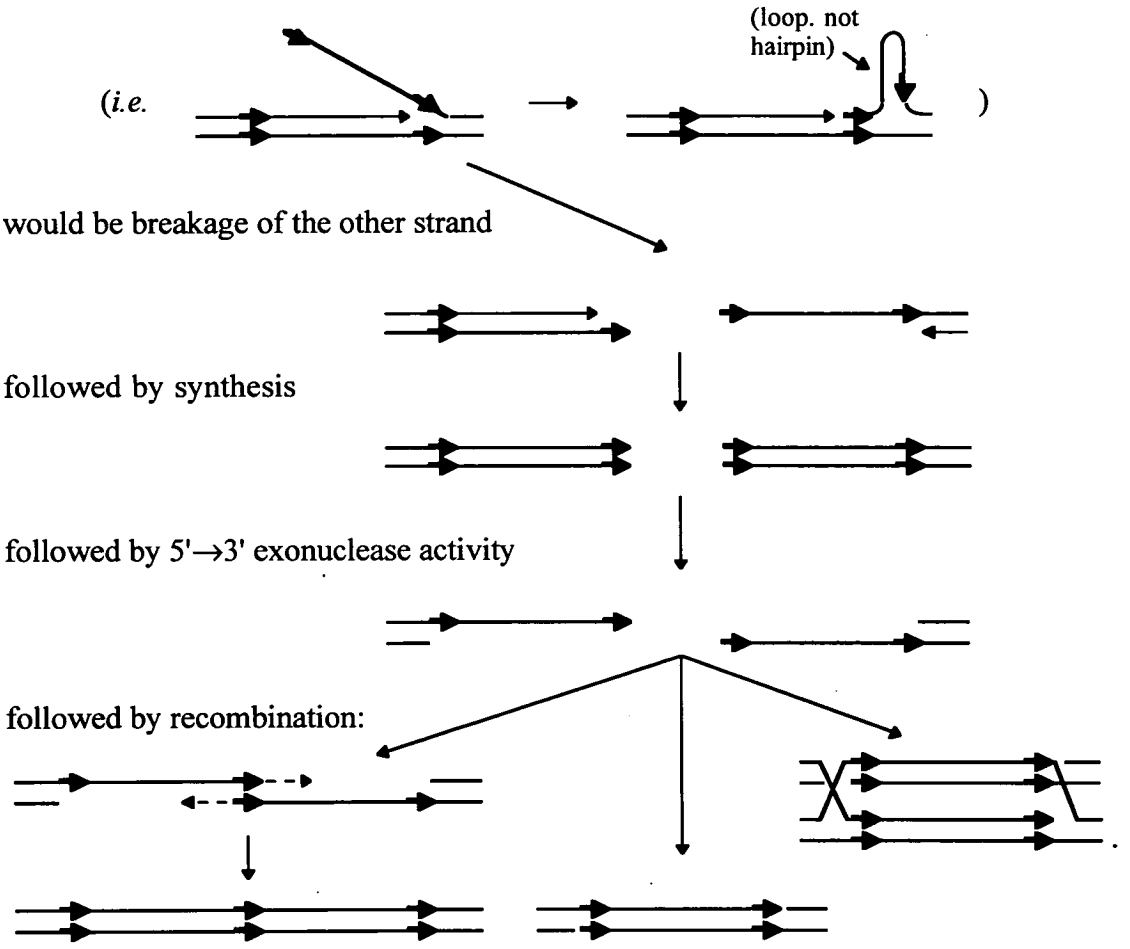
4. Length increases instability and interruptions decrease it but what is the mechanism?

Hypotheses abound. In the first paper to come from this project (Darlow & Leach, 1995), we addressed the question of how results with only a few repeat units might relate to instability in long arrays and suggested two possibilities. Firstly, small nucleating loop could constitute the rate-limiting step leading to the formation of a larger and more stable secondary structure which might be involved in exaggerated slippage. Secondly, as mentioned in Chapter 1, even a small secondary structure could result in a large change in number of repeats through cleavage and recombination in a long tract of repeats, especially if the latter is longer than the minimal length of homology required to initiate homologous recombination, the minimal efficient processing segment, about 200 and 300 bp in mammals. Since partial sequence divergence has been found to inhibit recombination severely in mammalian cells (Waldman & Liskay, 1987), this hypothesis could accommodate the suppression of instability by imperfect repeats.

The sliding Okazaki fragment hypothesis of Richards & Sutherland (1994) Sutherland & Richards (1998) discussed above was put forward to explain why massive expansions occur beyond a threshold number of repeat units. It would certainly be inhibited by variant repeats but it has another problem that was discussed.

The Okazaki fragment flap hypothesis of Gordenin *et al.* (1997) also discussed above is very plausible. The folding of a flap to form a hairpin and subsequent ligation of its 5' end to the next Okazaki fragment might only be expected to cause fairly small expansions unless flaps are usually large, which would be very wasteful, but there is the recombination possibility. Tishkoff (1997) examined the mutations that occurred in *RAD27* mutants of *S. cerevisiae* and found duplications of up to 108 bp. These were duplications of random sequence between two short direct repeats of 3 - 12 nt that were not always perfect with respect to one another, and one copy of the repeated sequence was included in the duplication. They then tested for

recombination and found that there was a 25% increase in mitotic recombination. They suggested that an alternative to pairing of the 5' short repeat in the flap with the 3' one in the template

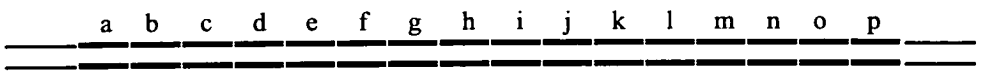


Clearly with a tract of trinucleotide or other tandem repeats the possibilities would be much wider than with only two repeats separated by unique sequence.

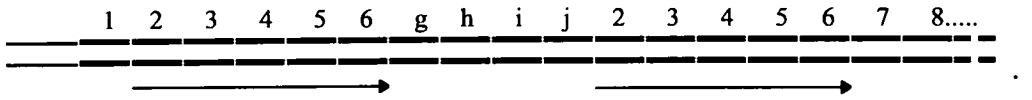
In Chapter 1 it was mentioned that Jeffreys *et al.* (1994) and Buard & Vergnaud (1994) had observed that new minisatellite alleles had been produced by complex recombination events and that the latter authors had suggested that these might come about by double-strand breaks that derived from staggered single-strand nicks, several repeat units apart, followed by the separation of the ends and repair by recombination with the other allele or sister chromatid. This could result, amongst other possibilities, in a new allele containing direct repeats of the chain of repeat units

from the overhang separated by copies of a row of repeat units from the other allele, *i.e.* recombination repair of this

with this



could yield this



Buard & Vergnaud (1994) did not draw the strand interactions but did draw a heteroduplex intermediate and four possible products. They said that either both gaps would be filled by strand transfer from the donor or one could be so filled and the other filled by synthesis. More recently, Buard & Jeffreys (1997) have suggested that the two free ends might invade the intact duplex, extend on their respective templates, and then come out again and reanneal with one-another in a new position, rather like the bottom left diagram on p. 309 but with a little more overlap, and then fill in by synthesis.

Whether this type of expansion mechanism applies to trinucleotide repeats is unknown. Neither paper (Buard & Vergnaud, 1994; Buard & Jeffreys, 1997) suggests how staggered nicks in the repeat tract might be generated. In a tract of d(CAG)·d(CTG) or d(CGG)·d(CCG) repeats it might possibly be by the cleavage of hairpins or other unusual secondary structure formed at different positions on the two strands. Sarkar *et al.* (1998) suggested that it might result from stalling of replication forks at secondary structure and in their model there were 5' overhangs rather than 3' ones, and their ends were annealed and the gaps filled in by synthesis.

Numerous further papers have come out of Jeffreys' laboratory since the above-mentioned ones. Jeffreys and Neumann (1997) examined somatic mutants of

the human 29-bp-unit minisatellite MS32 (mentioned in Chapter 1) by looking at DNA of blood leukocytes from two individuals by single-molecule PCR. They found that these consisted of deletions or duplications of blocks of repeat units located at random along the repeat tract. This pattern was quite different from the complex rearrangements seen in sperm DNA. The authors concluded that conversion-based mutation is germline-specific, most likely meiotic, and that somatic mutation occurs by a different mechanism involving replication slippage or unequal sister-chromatid exchange. They noted that, unlike the pattern in trinucleotide repeat expansions, the mutations did not cluster at homogeneous runs of identical repeats.

Jeffreys *et al.* (1998) examined conversion events and inter-allelic much rarer equal and unequal crossovers (with exchange of flanking markers) at MS32 and MS31 (a 20-bp-unit minisatellite also mentioned in Chapter 1) by single-sperm analysis. Both conversion and crossovers showed polarity and both were 'suppressed' in an unusually stable allele and the authors concluded that the two types of event occur by a common mechanism.

Buard *et al.* (1998) investigated the rates and types of mutation in a wide range of different alleles of the human 37-43-bp-unit minisatellite CEB1 (also mentioned in Chapter 1). They found that, as with trinucleotide repeat alleles, mutation rates varied greatly between alleles, by up to three orders of magnitude in this case. More interestingly, they found that different types of mutation showed different behaviour. Intra-allelic rearrangements increased with array size and here they did tend to cluster in homogeneous segments of alleles, both features seen in trinucleotide instability. In contrast, inter-allelic rearrangements occurred at a relatively constant rate with respect to array length and showed a mild polarity towards one end of the tract. It is still unknown why very unstable minisatellites with many different variant repeats should be able to recombine while a few interruptions by variant repeats in trinucleotide repeat tracts make the latter very much more stable but perhaps these findings suggest that much of trinucleotide

repeat expansion occurs by some kind of slippage mechanism, which has remained the popular choice from the start.

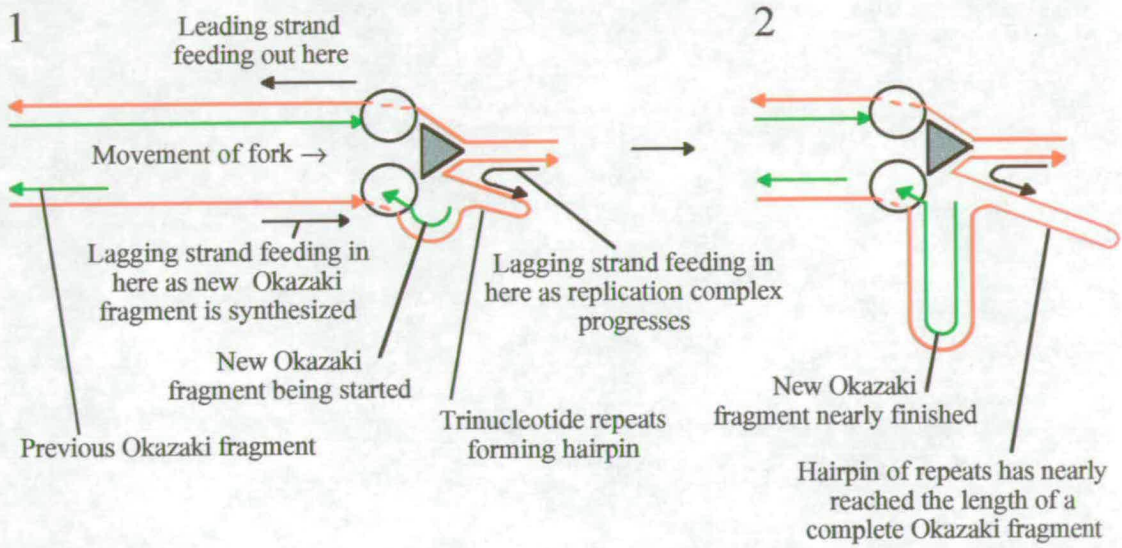
Harvey (1997) considered the energy involved in slippage. A little bulge in the DNA formed of one trinucleotide looped out would have an enthalpic cost. This would remain about the same for any length of tract but there would be an entropic advantage due to the number of different places where the bulge could be (which could be in any frame) and this would increase with the length of the tract. This, Harvey suggests, provides a simple thermodynamic basis for dynamic mutation. When there are few repeats, slips are rare and the repair mechanisms are able to correct them. Longer tracts are prone to more frequent slips that occasionally escape repair, leading to gradual expansion, and even longer tracts have high frequencies of slippage, overwhelming the repair systems, and expansion is rapid.

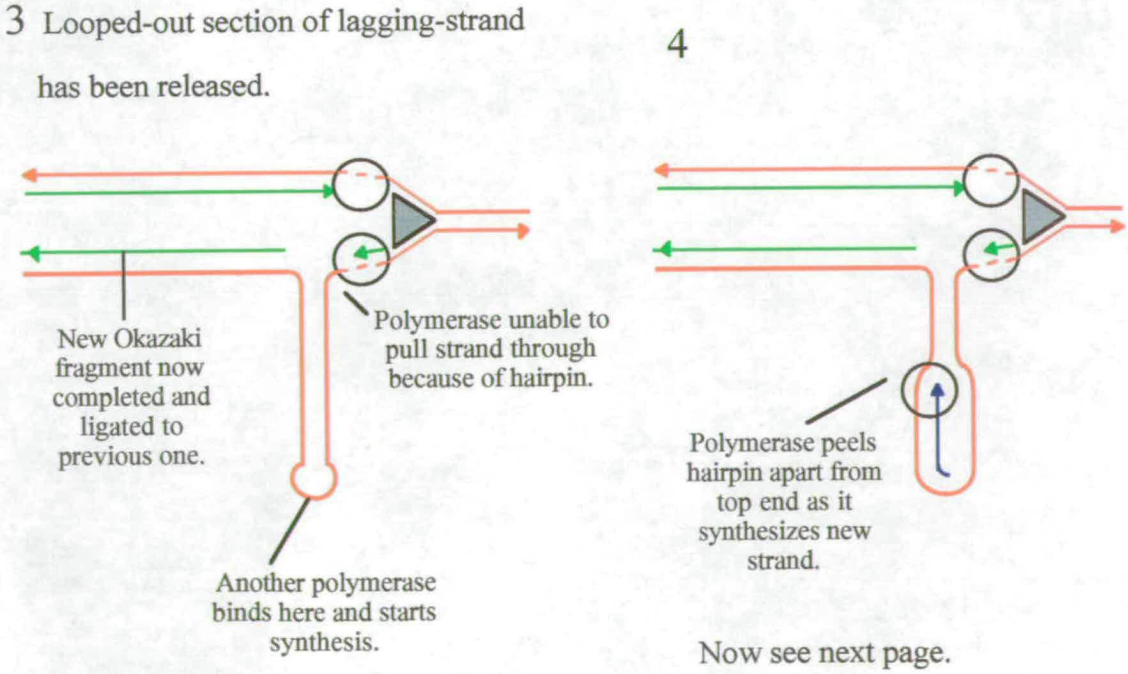
My first thought on this is: is it possible for repair to be locally overwhelmed but unaffected in the rest of the cell so that other loci are unaffected? Harvey (1997) makes no reference to the fact that his thermodynamic argument would apply equally to any trinucleotide repeat tract and cannot explain why some are prone to massive expansion and others not. He does go on to discuss hairpins, though he fails to acknowledge that only some tracts can make them. He believes that entropic considerations favour bulges over hairpins. If multiple slips cause multiple bulges, Harvey says, the second and subsequent slips have almost as much advantage as the first because they still have many possible locations. I dispute this because each can only be 3' to the previous one, considerably limiting the possibilities (unless the bulges move along). If, however, the first slip nucleates a hairpin that grows on subsequent slippages, each subsequent bulge has nowhere else to go but into the existing hairpin, so there is no increase in entropic advantage. The enthalpic cost, he says, is more difficult to evaluate. On the one hand hairpins appear to be favoured because the enthalpic cost of having a junction does not increase if other bulges are absorbed into the hairpin, but on the other hand the hairpins are not perfect Watson-Crick duplexes so there is an additional enthalpic cost as the hairpin stem grows.

It seems to me that, as we know that DNA containing hairpin junctions is very stable (Pearson & Sinden, 1996; Pearson *et al.*, 1998a), and since small bulges do appear to move (Schlötterer & Tautz, 1992) and so might move into a hairpin, this mechanism may be a possible one (*if* bulges could move into a hairpin before the repair system was able to catch them). It is likely that it would favour expansion over contraction. Furthermore, an expansion that results in a new allele several times the length of either allele of the parent must be produced by some kind of iterative process. It also seems likely that this mechanism could be inhibited by the presence of variant repeat units. Pearson *et al.* (1998b) studied the effect of interruptions on S-DNA formation by various lengths of repeat tracts from the *SCA1* and *FRAXA* loci. As in their previous studies, the DNA was cloned with flanking sequences in plasmids which were linearized, melted with alkali and reannealed by neutralization, and the products examined by electrophoresis and densitometry. The percentage and number of isomers (seen as different bands) of S-DNA increased with tract length. Tracts interrupted by variant repeats showed lower percentages of S-DNA with fewer isomers than pure tracts of the same lengths. The authors suggested that variant repeats might reduce genetic instability by any of three possible mechanisms: inhibition of inter-strand slippage, inhibition of intra-strand interactions, and reduction of the opportunity for slippage by limiting the position of slip-out nucleation.

Mismatch repair deficiency reveals that polymerase slippage is common but that its effects are relatively uncommon because the repair system usually corrects them. McMurray (1995) apparently ignored this when she remarked that "it is difficult to imagine how simple slippages can occur since the entire fork complex must slip if protein-protein contacts are not broken." (Presumably she was only referring to large slippages.) Because of this, she proposed a different model. In this, the single-stranded region of lagging strand, looped out from the polymerase complex, forms a single large hairpin. Single stranded binding protein is only able to bind this at the loop. Another polymerase molecule associates here and synthesizes new

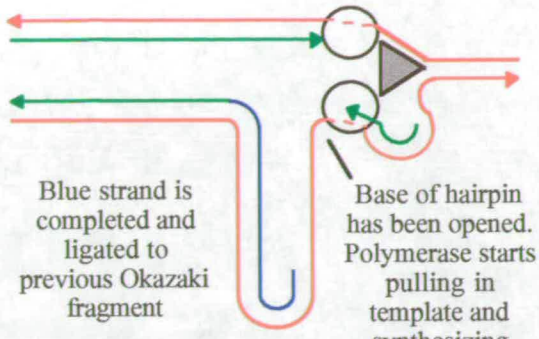
DNA up to the 5' end of the previous Okazaki fragment. This process disassembles the hairpin, allowing the polymerase complex at the fork to start a new Okazaki fragment. When this reaches the section of DNA that was synthesized on half of the hairpin it displaces exactly all of it (no more no less, apparently) and this loops out and is then ligated to the end of the new fragment. By this means, McMurray suggests, the tract would increase in length by 50%, which previous observations in her laboratory suggested is a common size of expansion in HD. This idea is illustrated below.





Now see next page.

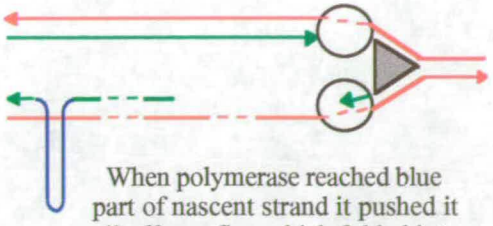
5



Blue strand is completed and ligated to previous Okazaki fragment

Base of hairpin has been opened. Polymerase starts pulling in template and synthesizing another Okazaki fragment.

6



When polymerase reached blue part of nascent strand it pushed it all off as a flap which folded into a hairpin and was ligated to the completed full-length Okazaki fragment.

If the repeat tract is long enough, the whole process could happen again.

McMurray (1995) did not draw all of these stages but her description in the text (and a figure in a later paper, Gacy & McMurray, 1998) makes it clear that this is what she meant. There are at least three objections to this model: (i) As discussed above, a hairpin cannot grow from only one side - unless single trinucleotide bulges can be fed into it and ripple up to the top - so the looped-out trinucleotide repeat single strand might be a series of small hairpins rather than one long one. This may not matter. (ii) Could RNA-primed DNA synthesis really be initiated in the hairpin loop by another complex? (iii) There is no problem with the polymerase complex pushing up a flap, or with that flap folding into a hairpin, as was later proposed by Gordenin *et al.* (1997). However, McMurray did not use the term 'flap' but somehow imagined that the polymerase would know when it had reached the preceding (green) Okazaki fragment though actually there would be no distinction between the blue and green parts of the strand. The length of the blue part of the strand is equal to half the length of repeats that formed a hairpin and if the repeats are on parts of more than one Okazaki fragment, this is supposed to happen each time.

As discussed in Chapter 4, Gacy & McMurray (1998) construed from their results that the reason for the length-dependence of trinucleotide repeat tracts for large expansions is that short hairpins reanneal with their complementary strand too quickly to cause trouble while long hairpins, though no more stable, take much longer

to reanneal with their complements and so can be trapped in the folded-out position by further replication. However, 'long', for them, meant 25 repeats, which is below the threshold for large expansions. Another claim that they made was that the thermal stabilities of $d(CTG)_n$ and $d(CAG)_n$ hairpins are very similar and that therefore the widely held notion that expansion occurs (on the lagging strand) when the nascent strand has the CTG repeats and contraction when the template has them because CTG hairpins are more stable than CAG ones should be revised. As the results quoted in Chapter 4 show, all investigators found CTG repeats to make the more stable hairpins but others found greater differences in T_m. Mitas and colleagues found a difference of 10°C (Yu *et al.*, 1995a,b) but only used 1 mM NaCl, and Petruska *et al.* (1996) found a difference of 4 - 5°C at 19 mM NaCl and 6°C at 170 mM NaCl, while Gacy & McMurray (1998) found only 1·3°C at 100 mM NaCl. Since Gacy & McMurray (1998) believe that there is little difference in thermal stability, they suggest that perhaps protein interactions may be better at preventing hairpin formation by CAG strands than by CTG ones.

The only protein suggested by name by Gacy & McMurray (1998) was single-stranded binding protein (SSB). It has been shown that deficiency of SSB in *E. coli* considerably increases the rate deletion of d(CAG)·d(CTG) repeats cloned in the orientation in which the deletion rate is usually the lower (CAG on the lagging-strand template) (Rosche *et al.*, 1996). This is good confirmation that instability results from secondary structure formation but does not answer the question of whether SSB could contribute to the differential stability of the structures formed by the two strands. However, as mentioned earlier, it has been found that MSH2 binds better to CAG hairpins than to CTG ones (Pearson *et al.*, 1997) so this might act as a protector, recognizing the mismatches and binding but unable to initiate repair. It is clear anyway that mismatches within hairpins are not corrected because there have been no reports of changes of CAG to CTG or *vice versa* within a single strand.

Though the two strands of d(CGG)·d(CCG) repeats form different structures which might intrinsically affect which orientation is more prone to expansion,

proteins may also be involved here. HMG box proteins form the most abundant class of non-histone chromosomal proteins and have been shown to interact strongly with branched or cis-platin reacted DNA, relative to normal duplexes. Since a trinucleotide repeat expansion is thought to involve unusual DNA structure, Zhao *et al.* (1996) decided to test for the binding of an HMG box protein. They annealed a constant amount of labelled $d(CCG)_{15}$ with various concentrations of unlabelled $d(CGG)_{10}$ in 100 mM NaCl, 50 mM tris·HCl at pH 7·5 and obtained five complexes separable by electrophoretic mobility. Three of these were susceptible to digestion by Exonuclease VII and were taken to have single-strand tails but the other two were not digested and were taken to be branched structures. Their nature was not speculated upon but one (or perhaps both) was presumably a Watson-Crick duplex with a $d(CGG)_5$ hairpin branching out. HMGb, an 81-residue polypeptide containing the second HMG box motif of rat HMG1, bound to one of these complexes causing a gel-mobility shift. The authors took this to be evidence that the DNA formed a branched structure and suggested that HMG box proteins might stabilize slipped structures. Perhaps therefore these could also be candidates for distinguishing CAG from CTG, if any is needed.

Another protein which appears to bind selectively to d(CCG) hairpins is human DNA-(cytosine-5)-methyltransferase. The ideas of S.S. Smith and colleagues on the methylation of d(CGG)·d(CCG) repeat tracts *via* methylation of d(CCG) hairpins has been mentioned earlier. More recently (Kho *et al.*, 1998) they have found severe product inhibition of the enzyme by the methylated d(CCG) hairpin and demonstrated a DNA protein complex by gel-retardation. From this they have concluded that the enzyme becomes tightly bound to the hairpin and 'stalls', because the ^{m}C·C-mismatch is similar to the transition state in a Watson-Crick duplex, and that this may be involved in the expansion mechanism, probably slippage. Their thoughts on d(CAG) hairpins will be mentioned later.

Returning to Gacy & McMurray (1998), the point remaining to be discussed is their contention that the reason why perfect palindromic sequences are less liable

to expansion mutations than trinucleotide repeat tracts of the same length is that though both form hairpins that take a long time to reanneal with their complements, palindromes rarely form hairpins in the first place but trinucleotide repeats do so frequently. The fact that long palindromes are either deleted or inviable in wild-type *E. coli*, depending upon their length, indicates that palindromes are very liable to hairpin formation. What Gacy & McMurray fail to mention is that, unlike trinucleotide repeat tracts, palindromes have exactly the same sequence on both strands so whenever the nascent strand is liable to hairpin formation so will the template strand be also, and exactly as stably.

Part of the reason for our susceptibility to repeat expansion may result from our tolerance of palindromic, and hence quasi-palindromic, sequences. Sarkar *et al.* (1998) have obtained large expansions on cloning long d(CAG)·d(CTG) tracts in *E. coli* when orientated (as usual) so that the CTG repeats would be on the new strand on lagging-strand replication. They used SURE cells (Stratagene) which have been used by others for cloning trinucleotide repeats (Shimizu *et al.*, 1996) without such success in achieving expansions. The additional factor introduced by Sarkar *et al.* (1998) was growing the bacteria at 25°C or below in order to increase the stability of unusual secondary structures. They measured the size of Okazaki fragments in SURE cells by pulse labelling and found them to range from 0·9 to 1·2 kb and found that the frequency of expansion increased with length of the repeat tract beyond 330 repeats (*i.e.* about 1 kb) which corresponds well with this. Not surprisingly they also found that the proportion of expanded products increased with the number of generations. SURE cells have a mutation in *sbcC*, such as allowed long palindromes to be propagated for the work reported in this thesis. They also have mutations in many other genes, including *recB* and *recJ*, but Sarkar *et al.* (1998) showed that lack of functional SbcC was important for expansion by producing SURE *sbcC*$^+$ cells. In these cells replication of plasmids with 2 or 20 repeats was unaffected but the yield of plasmids bearing 120 or more repeats was apparently virtually eliminated and those plasmids that were recovered were said to have severely deleted tracts (500

down to <100 repeats). Their preferred model for large expansions was the updated (Sutherland *et al.*, 1998) model of the Okazaki fragment sliding in a 5' direction and producing a very large flap, which they drew not as a hairpin but as a branched structure, the end of which would be ligated to the following fragment. This would be immune to attack by FEN1 but in bacteria would be cleaved by SbcCD and degraded. The authors pointed out that it will be interesting to discover the possible differences in substrate recognition by SbcCD and the human protein RAD50/ MRE11 which shares homology with it, quoting Sharples & Leach (1995). Actually the latter authors used the sequences of the yeast proteins, but human Rad50 and Mre11 have now been found. It has been shown that Mre11 does cleave hairpins that contain mismatches (Paull & Gellert, 1998) so the mystery deepens.

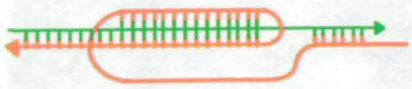
Another point that Sarkar *et al.* (1998) made is that the human Okazaki fragment length has been estimated at 100 - 200 bp, but the bacterial one as 1,000 - 2,000 bp and that this may be the reason why human repeat tracts become liable to expand at lengths (they cite 50 repeats) at which they would expect bacterial ones to remain stable. (They do not tackle the fact that repeats actually tend to delete in bacteria.) Though favouring the sliding Okazaki fragment model, they did not rule out the possibility of a recombination model that was mentioned earlier. The recent discovery that instability of the *FRA10B* locus seems to occur at about 75 repeat units, where the average unit length is not 3 bp but 42 bp, suggests however that stability may depend upon the number of copies of the repeat unit, rather than on the length of the repeat tract measured in base pairs. It may also be that the threshold is different depending upon the sequence of the repeat unit; AT-rich repeats might need more copies for instability than CG-rich ones, as suggested by Gacy *et al.* (1995) in relation to the dinucleotide repeats $(AT)_n$.

Lastly, on this question, any proposed mechanism must be able to explain expansion of d(GAA)·d(TTC) repeat tracts. I have not discussed in detail the possible structures formed by this repeat as my work has not been involved with triplexes. The two strands of this repeat are all-purine (R) and all-pyrimidine (Y).

Triplexes could either form with one pyrimidine and two purine strands, Y·R·R, or with one purine and two pyrimidine strands, Y·R·Y, and in either case, the two strands the same could either be parallel or antiparallel. From duplex DNA a triplex can be formed by one strand doubling back, *i.e.* going backwards and then forwards again to continue in its original direction, and depending upon which parts of the folding strand are paired with the non-folding one, different types of triplex may be formed. The strands can always be aligned so as to have C·G·G and T·A·A or C·G·C and T·A·T triads without any necessity for mispairing of A with G *etc.*.

By melting profiles (*v.* optical density) of different oligonucleotides, Gacy *et al.* (1998) showed that both Y·R·R and Y·R·Y triplexes would form but only with the third strand parallel to the d(GAA) strand of the Watson-Crick duplex (thus in the Y·R·Y triplex the two contributing parts of the d(TTC) strand are antiparallel to one-another). The authors oddly concluded that in the genome the Y·R·R one would have difficulty in forming because of the way it would have to fold and that therefore only the Y·R·Y one was relevant, whereas in fact either could be formed just as easily as far as topology of strands is concerned. However, the melting profile clearly showed that the Y·R·R one was less stable. C·G·C triads require protonation of one of the cytosine residues whereas C·G·G triads do not require cytosine protonation. Gacy *et al.* (1998) showed that replication of a $d(GAA)_n$ strand on $d(CTT)_n$ templates, of various lengths up to 250 repeats, is impeded at pH 8, severely impeded or blocked at pH 6 or pH 7 depending on the length. The only synthesis of a d(TTC) strand on a d(GAA) template tried was with 28 repeats and this caused a 5% blockage at 14 repeats at pH 7·1. At the same pH and length with the other template the blockage was 40%. Thus the Y·R·Y triplex is definitely the more important. Mariappan *et al.* (1999) have shown by NMR that in a 1:1 mixture of $d(GAA)_3$ and $d(TTC)_3$ a finite population of triplex exists with C^+·G bonds so this confirms that the Y·R·Y triplex is thermodynamically preferred. Since it is the d(TTC) strand that is most likely to double back, if this strand is the template in replication a deletion will result while if it

is the nascent strand an expansion will occur. The looping required is shown below; green represents the d(GAA) strand.



Deletion

Expansion

Gacy *et al.* (1998) studied the transmission of long repeat alleles to determine the mechanism of expansion. Friedreich's ataxia is autosomal recessive. First they looked at the lengths of the alleles in the offspring of two heterozygous parents. Alleles of normal length were inherited stably whether the offspring were homozygous normal or heterozygous and there was no difference in the instability of the long allele whether the offspring were heterozygous or homozygous (affected). This showed that expansion did not depend upon allele interaction in the offspring. To determine whether allele interaction occurred in the parents (meiosis) the authors compared the frequency of instability between heterozygous offspring of either one homozygous affected and one normal parent or two heterozygous parents. Neither the size nor the frequency of repeat length changes was different in the two groups. From this the authors concluded that instability does not require interaction of alleles and therefore must occur by an intra-allelic mechanism. This of course does not exclude recombination following cleavage at unusual secondary structure during replication.

In a review, Mitas (1997) announced that he and his co-workers had experimental data indicating that $d(GAA)_{15}$ adopts an unconventional hairpin with G·A base-pairs and suggested that despite others' suggestions that d(GAA)·d(TTC) tracts might expand by a different mechanism from d(CAG)·d(CTG) and d(CGG)·d(CCG), only those trinucleotide repeat sequences that could form hairpins might be able to expand. Mariappan *et al.* (1999) claim that, contrary to the claim of Mitas and colleagues, they have shown that d(GAA)·d(TTC) repeats do not form

hairpins. Actually they have done nothing of the kind. They have performed NMR studies with the oligonucleotides d[$(GAA)_2T_4(TTC)_2N_n(CTT)_2$] where N_n = TTTT or TTCTT or TT. These oligonucleotides were specifically designed to investigate the particular type of triplex believed to form and could not possibly have formed a $d(GAA)_n$ hairpin. Now, Mitas and colleagues (Suen *et al.*, 1999) have presented their results. They have investigated $d(GAA)_{15}$ by circular dichroism, ultraviolet absorbance melting profile, base-modification and cleavage, and enzymic cleavage, and have concluded that in 50 mM NaCl at pH 7·5 the oligonucleotide forms a hairpin at 5°C, and an unknown but partially-base-paired structure at 37°C. *If* they are proved wrong, this would throw doubt on their other work such as that on the $d(CCG)_{15}$ hairpin. However, as mentioned before, Lee (1990) studied the effect of mono- and di-valent cations on $d(GAA)_n$, using ultraviolet absorbance melting profiles and concluded that it will form a quadruplex.

5. Questions of flanking sequences and polarity of expansion

In *in vivo* experiments on the stability of trinucleotide repeat tracts in *E. coli* and *Saccharomyces*, flanking sequences from the genes are nearly always included. No doubt this is partly because it is easier and cheaper to clone an authentic repeat sequence than to make a long one artificially, but surely there is also be an element of "better include the flanking sequence just in case it makes a difference". Several events and observations have led to this caution.

Before the dawning of the age of repeat expansion disorders, it was noticed that tandem repeat tracts that differed by a single base-pair in the repeat unit were often found in close proximity and were often contiguous (Levinson & Gutman, 1987 and refs therein). It was pointed out that this could readily be understood as the consequence of multiple small slippage events happening before and after base-substitution events. It was mentioned in Chapter 1 that Stallings (1994) found in a search of the human, mouse and rat genome databases (slight as they were in 1994 compared with today) seven examples of genes containing two repeat tracts that were

very close together and/or differed from one-another by only one base-pair. In the original report giving the sequence of the *IT15* gene of HD (Macdonald *et al.*, 1993) it was noted that there was a CCG repeat tract just downstream of the CAG one. The sequence on the coding strand was $(CAG)_n$,CCA,CAG,CCG,CCA,$(CCG)_7$. The number of copies of CCG was 7 on all three of the clones that were sequenced but the authors said that they could not exclude the possibility that this was also variable. The codons, with the two on either side, code for a run of 11 proline residues Actually the published sequence showed that the sequence beyond the 7 CCGs was quite rich in CCG and CAG trinucleotides, many codons coding for proline, including two more runs of 3 CCG codons.

As was mentioned in Chapter 1, Rubinsztein *et al.* (1993a,b) discovered that there was polymorphism in the CCG-rich region which could interfere with the estimates of lengths of the CAG tract by those laboratories whose PCR primers spanned the whole region. Rubinsztein *et al.* (1993a), using primers that just spanned the CCG-rich region, found four alleles of 176, 179, 182 and 185 bp. The frequencies of these alleles in 42 controls were 43, 4, 7 and 30 respectively, but in 44 HD patients 69, 0, 2 and 17. In the heterozygotes the phases were not determined (*i.e.* which CCG repeat was on the same chromosome as the expanded CAG repeat). Andrew *et al.* (1994) looked only at the first pure CCG stretch and found five alleles in controls. 66·8% of alleles had 7 copies, 29·8% had 10 copies and the remainder had 9, 11, or 12 copies. In HD patients they did examine phase, where necessary. Of 113 HD patients, the expanded CAG tracts of 105 (93%) were associated with 7 copies of CCG and the other 8 (7%) had 10 copies. Those with 7 all had 7 on the normal allele as well and of those with 10, three had 10 in the normal allele and the rest 7, the rarer alleles being absent.

In both surveys the differences between HD patients and controls were significant but no conclusions were drawn about this, the prime concern being avoidance of misdiagnosis of HD susceptibility. However, Barron *et al.* (1994) had wider interests. They used the same primers that Rubinsztein *et al.* (1993a) had

used, and found five alleles, the additional one being 170 bp. Of 568 control alleles 61% had the 176 bp allele, 31% the 185 bp, 5·5% the 182 bp, 2·4% the 179 and a single allele was 170 bp. In contrast, of 131 HD alleles, 130 were 176 bp and one was 179. The authors noted that the senior affected members of their patient families came from various parts of the British Isles and from other countries. They also noted that they were heterogeneous for the presence of a rare *Alu* insertion in an intron of a closely-linked gene. In addition they found a haplotype difference between patients from the east and west coasts of Scotland. From all this they concluded that though they could not completely rule out a founder effect they believed that the CCG-rich sequence might also be involved in the mechanism of disease. They suggested that perhaps the longer stretch of proline coded for by the longer common allele of the general population (185 bp) might be incompatible or perhaps lethal when inherited with the expanded CAG repeat that causes HD.

Barron *et al.* (1994) made no suggestion that the d(CCG)·d(CGG) tract might be involved mechanistically in the expansion of the d(CAG)·d(CTG) tract. After all, it was the shorter of the common CCG alleles that was associated with the expanded CAG. However, the Editor of Nature did suggest that the shorter common CCG allele might be necessary though not sufficient for the expansion of the CAG repeat (Maddox, 1994) and, as mentioned in Chapter 4, Gacy *et al.* (1995) used a modified RNA folding program to predict the possible secondary structures of the coding strand repeat sequences responsible for HD, SCA1 and DM and FRAXA and concluded that the flanking sequences were involved in the structures.

For the HD repeat, the prediction of Gacy *et al.* (1995) was that the CCG repeats pair with the CAG repeats at the base of a hairpin involving the whole tract. This makes C-A (and A·A) mispairs and Smith & Baker (1997) have reproduced this *in vitro* using a 48-mer oligonucleotide, d[$(CAG)_9$CAA,CAG,CCG,CCA$(CCG)_3$]. Electrophoresis showed that the oligonucleotide formed secondary structure (presumed to be a hairpin) and rapid methylation by the human DNA-(cytosine-5)-methyltransferase showed that this had the correct recognition site. From the finding

of product inhibition of methylation the authors concluded that the enzyme 'stalls' and remains bound to the hairpin. They proposed that since the enzyme has the capacity both for *de novo* and methyl-directed methylation it will probe the nascent strand or transcriptionally active DNA until it encounters an 'SSC' (slipped-strand conformer) where it will methylate cytosines and eventually stall. Thereby it will play an active rôle in biological processes that must distinguish between DNA in Watson-Crick-paired conformation and unusual conformations. They said that this might include DNA repair, recombination, and the maintenance of the differentiated state, but did not actually say that the enzyme might be involved in repeat expansion in this paper.

In myotonic dystrophy there is also belief that a sequence outside the repeat might be involved. Mahadevan *et al.* (1993) published the sequence of an insertion/deletion mutation in the *DMPK* gene, a deletion between direct repeats in the second and fifth of five consecutive *Alu* elements in an intron of the gene. These investigators also found that all of their DM cases had the larger *Alu* allele on the chromosome with the CTG expansion and they concluded that either the expansion mutation was quite ancient and occurred only a few times on the larger *Alu* allele or that the larger allele somehow predisposed to the expansion mutation. It was mentioned in Chapter 1, in the context of illustration that there are both large and small changes in repeat tract length, that Imbert *et al.* (1993) found that there was complete linkage disequilibrium between the insertion allele and repeat tracts with 5 CTG repeat units and with ones with 19 - 30 units and all the disease-range tracts, while many of the DM alleles with 6 - 17 units were on chromosomes with the other allele of the dimorphic *Alu* marker. Imbert *et al.* (1993) concluded that there had long ago been a single, or a very few, mutation(s) taking 5 repeats into the 19 - 30 repeat range and that these latter marginally-stable non-pathogenic alleles formed a pool from which mutations would from time to time occur, first into the 30 - 50 range (in which a few alleles were known) and from there into the disease range. Later, Neville *et al.* (1994) presented a high resolution genetic analysis of the locus using nine

polymorphisms spanning 30 kb within and immediately flanking the DM kinase gene and supported this hypothesis.

Subsequently the *DMPK* gene has been examined in a large number of normal populations and it has been found that in some populations some individuals have the *Alu* deletion associated with $(CTG)_5$ and $(CTG)_{>19}$ (e.g. Rubinsztein *et al.*, 1994; Zerylnick *et al.*, 1995; Tishkoff *et al.*, 1998) and a single case of DM has been reported in a sub-Saharan African and this person has the expansion on an *Alu* deletion allele (Krahe *et al.*, 1995). However, though two of the earliest reports of the DM expansion (Harley *et al.*, 1992; Yamagata *et al.*, 1992), which were published before the insertion/deletion mutation had been characterized, appear to show that not all expanded tracts were on the same background, it is claimed (Tishkoff *et al.*, 1998) that no non-African DM case has ever been found on a background other than (+++) for the *Alu* and two other markers and that the possibility that this allele predisposes to expansion cannot be excluded.

In the *MJD1* gene (of Machado-Joseph Disease, MJD, SCA3) the codon immediately following the $(CAG)_n$ may be either CGG or GGG, *i.e.* the repeat tract is followed by C or G. Limprasert *et al.* (1996), in a study including people from five racial groups, found that the C variant was present in all chromosomes with the expansion and 54·5% of chromosomes with 27 - 40 repeats but in none of those with less than 20 repeats. They also surveyed three other primates and found that chimpanzees also have the dimorphism and the C variant was present on both alleles with their largest CAG repeat number, 20, as it was on nearly all human alleles with 20 or 21 repeats. In macaques and mangabeys it was always G and they had CAG repeat ranges of 13 - 14 and 16 only respectively. The authors concluded that the C variant may influence the CAG repeat stability but that since it is very rare in alleles with 22 - 26 repeats there must another factor in addition. Matsumura *et al.* (1996) presented similar findings from a Japanese population. They remarked that the nucleotide following the CAG repeat tract in SBMA, HD, SCA1 and DRPLA is also C and that "the mechanism by which the $(CAG)_nC$ configuration would be prone to

instability is unknown". A large collaborative survey of several ethnic groups (Igarashi *et al.*, 1996) found that the difference in intergenerational size change in the repeat from affected parent to child was not significantly greater if the expanded repeat ended with the C or the G variant. However, the frequency of changes of ±(>2) (mainly expansions) was very significantly greater when the C variant was on the expanded allele and the G on the normal allele than when both were C or both were G. (There were too few cases in which there was a G with the expanded tract and a C on the normal allele for statistical purposes.) A study of sperm from MJD patients (Takiyama *et al.*, 1997) confirmed increased instability of expanded $(CAG)_nC$/normal $(CAG)_nG$ over the homozygotes for the dimorphism, but surprisingly deletions outnumbered expansions in the sperm.

Pearson *et al.* (1998b) found that 19·3% of the DNA containing a pure $d[(CAG)\cdot(CTG)]_{30}$ tract cloned with SCA1 flanking sequences (107 bp on one side and 358 on the other) formed S-DNA whereas in their previous study (Pearson *et al.*, 1997) they had found 39% S-DNA with the same length tract cloned with DM flanking sequences (59 bp on one side and 54 on the other). They concluded that the flanking sequences can influence the percentage and pattern of S-DNA products formed but did not venture to suggest how this should be.

Jeffreys and colleagues have been trying to find flanking sequence elements responsible for repeat instability in the human minisatellites, MS31A, MS32 and CEB1, all mentioned previously. In Chapter 1 it was mentioned that Jeffreys *et al.* (1994) suggested that mutational polarity of the recombination events they detected implied that mutation was modulated by element(s) outside the array and proposed activation of the recipient locus by introduction of a double-strand break by a protein binding to a mutation-initiator sequence near the end of the repeat array at which the expansions occurred, with the position of the break controlled by the initiator. Monckton *et al.* (1994) found a C/G base-substitution dimorphism 48 bp away from the unstable end of MS32. Repeat tracts on the same strand as the C variant were found to be unusually stable but they did act as conversion donors of stretches of

repeats to mutant alleles formed on the G strand. This led Igarashi *et al.* (1996) to suggest that their findings *re* instability and the C/G dimorphism associated with the MJD (SCA1) repeat might indicate that gene conversion might be involved in the instability in this disorder. However, remember that Buard & Vergnaud (1994) noticed that the polarity in the interallelic exchanges observed by themselves and Jeffreys *et al.* (1994) could be accounted for by exact alignment of the same ends of the two alleles without the need to invoke a *cis*-acting element. Now Jeffreys and colleagues (Murray *et al.*, 1999) have announced the results of comparative sequence analysis performed to search for common flanking elements associated with MS31A, MS32 and CEB1. All three minisatellites were found to be located in GC-rich DNA abundant in dispersed and tandem repetitive elements but there were no significant sequence similarities between the different loci upstream of the unstable end of the repeat arrays and no consistent patterns of thermal stability or DNA secondary structure were shared by DNA flanking the repeats. The authors conclude that recombinational activity is not controlled by primary or secondary characteristics of the DNA sequence flanking the repeat array and is not obviously associated with gene promoters as seen in yeast.

Willems (1994) remarked that many of the trinucleotides that closely followed the known CAG repeat tracts differed from CAG by only one nucleotide. This was not an entirely new observation; simple-sequence repeats tend to be embedded in similar sequence (for examples see Levinson & Gutman, 1987). In the spring of 1995 (J.M. Darlow, Ph.D. First Year Report, unpublished), I looked at the sequences flanking the repeat tracts then known to be associated with inherited disorders and came up with the following hypothesis. "In both CAG and CGG repeat tracts single base-change mutations may occur at any point. Expansion tends to occur most frequently in the longest run of uninterrupted repeats and so the variant trinucleotides tend to get pushed, so to speak, to the peripheries and may go in either direction depending upon where mutation was in the tract. Once marginalized, the variant trinucleotides, and any portion of the original tract cut off,

will continue to mutate further (subject of course to the constraints of any selection). Thus the repeat tract is continually expanding in the centre and degenerating at the edges. The presumably recent expansion of pure CGG repeats right at one edge of an older degenerated tract in *FRAXF* may not detract from this hypothesis. Back mutations or loss of variant trinucleotides may relink separated parts of pure repeat. Apparent polarity of expansion of the whole tract results from the longest pure tract happening to be at one side and does not necessarily reflect any real polarity in the expansion of the pure tract itself. Such polarity, if it occurs, is undetectable by sequence comparisons but, as the *FRAXA* data suggests polarity in the loss of AGG trinucleotides, there may also be a polarity of mutation of pure tracts." By 'expansion' in this context, I meant the slow increase in size of repeat tracts by small slippages. Since repeat tracts tend to be flanked by degenerate repeats, the observation of Matsumura *et al.* (1996) that four expanding CAG repeat tracts are followed by C is not at all surprising and does not necessarily have any significance.

More recently, Eichler *et al.* (1996) have carried out a haplotype and interruption pattern analysis of normal and premutation *FRAXA* repeat tracts and have shown that there are two different mutational pathways that have originated fragile X syndrome. There can be up to four AGG interruptions in the CGG strand but most commonly two. Haplotypes were specified by three polymorphic markers, 7, 11 and 150 kb from the repeat tract. On some haplotypes, rare in the general population there has been loss of the 3' interruption, probably on many separate occasions, followed by probably fairly rapid increase of the longest pure part of the tract by slippage to reach the unstable range, but on one haplotype two interruptions have been retained and there appears to have been a slow steady increase in length of the most 3' part tract to generate a large pool of alleles predisposed to disease mutation. From this the authors concluded that there may be haplotype-specific influences in both the loss and maintenance of interruptions. Polarity is seen in two aspects. Firstly, though all pure stretches between the interruptions have sometimes changed in size only the most 3' stretch (on the CGG strand) has ever reached an

unstable length. Secondly there has been a considerable bias among normal alleles towards loss of the most 3' of the interruptions. Once a long pure tract has arisen, that is where the large expansions are going to be, so it may only be small slippages and losses, mutations or conversions of interrupting repeats that are actually polar.

I repeat that this has only been a review of a small selection of all the many papers on the mechanisms of repeat instability and it is not possible here to be comprehensive. I can only hope that it is a reasonably good selection. It seems that though recombination undoubtedly contributes to repeat tract instability the major factor is more likely to prove to be some kind of slippage mechanism. Since the sex of the parent often has a major effect on the degree of stability and massive expansions seem only to occur in gametogenesis or in the first few cell divisions of the embryo there are undoubtedly *trans*-acting factors that influence stability apart from the normal mechanisms of replication, recombination and repair. Human cDNA transgene constructs with up to 82 d(CAG)·d(CTG) repeats have remained stable over multiple generations in mice while genomic fragments with as few as 55 of the same repeats have proved unstable (Longo & Massa, 1997 and refs therein). However, cloning in genomic sequence is not always successful. Long pure tracts of up to 97 d(CGG)·d(CCG) repeats within the first exon of the human *FMR1* gene have been found to be stable apart from increases and decreases of one or two units (Lavedan *et al.*, 1998). Much remains to be done.

Bibliography

Aaltonen, L. A., Peltomäki, P., Leach, F. S., Sistonen, P., Pylkkänen, L., Mecklin, J.-P., Järvinen, H., Powell, S. M., Jen, J., Hamilton, S. R., Petersen, G. M., Kinzler, K. W., Vogelstein, B. & de la Chapelle, A. (1993). **Clues to the pathogenesis of familial colorectal cancer.** *Science* **260**, 812-816.

Acevedo, O. L., Dickinson, L. A., Macke, T. J. & Thomas, C. A., Jr. (1991). **The coherence of synthetic telomeres.** *Nucl. Acids Res.* **19**, 3,409-3,419.

Allers, T. (1993). **Detection of cruciform DNA *in vivo*.** Unpublished thesis, University of Edinburgh.

Andrew, S. E., Goldberg, Y. P., Theilmann, J., Zeisler, J. & Hayden, M. R. (1994). **A CCG repeat polymorphism adjacent to the CAG repeat in the Huntington disease gene: implications for diagnostic accuracy and predictive testing.** *Hum. Mol. Genet.* **3**, 65-67.

Ashizawa, T., Anvret, M., Baiget, M., Barceló, J. M., Brunner, H., Cobo, A. M., Dallapiccola, B., Fenwick, R. G. J., Grandell, U., Harley, H., Junien, C., Koch, M. C., Korneluk, R. G., Lavedan, C., Miki, T., Mulley, J. C., Munain, L. d., A. Novelli, G., Roses, A. D., Seltzer, W. K., Shaw, D. J., Smeets, H., Sutherland, G. R., Yamagata, H. & Harper, P. S. (1994). **Characteristics of intergenerational contractions of the CTG repeat in myotonic dystrophy.** *Am. J. Hum. Genet.* **54**, 414-423.

Ashley, C. T., Sutcliffe, J. S., Kunst, C. B., Leiner, H. A., Eichler, E. E., Nelson, D. L. & Warren, S. T. (1993). **Human and murine *FMR-1*: alternative splicing and translational initiation downstream of the CGG-repeat.** *Nature Genet.* **4**, 244-251.

Ashley, C. T., Jr. & Warren, S. T. (1995). **Trinucleotide repeat expansion and human disease.** *Annu. Rev. Genetics* **29**, 703-728.

Aslanidis, C., Jansen, G., Amemiya, C., Shutler, G., Mahadevan, M., Tsilfidis, C., Chen, C., Alleman, J., Wormskamp, N. G. M., Vooijs, M., Buxton, J., Johnson, K., Smeets, H. J. M., Lennon, G. G., Carrano, A. V., Korneluk, R. G., Weiringa, B. & de Jong, P. J. (1992). **Cloning of the essential myotonic dystrophy region and mapping of the putative defect.** *Nature* **355**, 548-551.

Ayares, D., Chekuri, L., Song, K.-Y. & Kucherlapati, R. (1986). **Sequence homology requirements for intermolecular recombination in mammalian cells.** *Proc. Natl. Acad. Sci. U.S.A.* **83**, 5,199-5,203.

Bacolla, A., Gellibolian, R., Shimizu, M., Amirhaeri, S., Kang, S., Ohshima, K., Larson, J. E., Harvey, S. C., Stollar, B. D. & Wells, R. D. (1997). **Flexible DNA: genetically unstable CTG·CAG and CGG·CCG from human hereditary neuromuscular disease genes.** *J. Biol. Chem.* **272**, 16,783-16,792.

Barnicoat, A. J., Wang, Q., Turk, J., Green, E., Mathew, C. G., Flynn, G., Buckle, V., Hirst, M., Davies, K. & Bobrow, M. (1997). **Clinical, cytogenetic, and molecular analysis of three families with FRAXE.** *J. Med. Genet.* **34**, 13-17.

Barron, L. H., Rae, A., Holloway, S., Brock, D. J. H. & Warner, J. P. (1994). **A single allele from the polymorphic CCG rich sequence immediately 3' to the unstable CAG trinucleotide in the IT15 cDNA shows almost complete disequilibrium with Huntington's disease chromosomes in the Scottish population.** *Hum. Mol. Genet.* **3**, 173-175.

Bates, G. & Lehrach, H. (1994). **Trinucleotide repeat expansions and human disease.** *BioEssays* **16**, 277-284.

Behn-Krappa, A. & Doerfler, W. (1994). **Enzymatic amplification of synthetic oligodeoxyribonucleotides: implications for triplet repeat expansions in the human genome.** *Human Mutation* **3**, 19-24.

Bell, G. I., Horita, S. & Karam, J. H. (1984). **A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus.** *Diabetes* **33**, 176-183.

Bell, G. I., Selby, M. J. & Rutter, W. J. (1982). **The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences.** *Nature* **295**, 31-35.

Bell, J. with clinical notes by J. P. Martin (1947). **Dystrophia myotonica and allied diseases.** *Treasury of Human Inheritance* **IV**, 343-410.

Bell, M. V., Hirst, M. C., Nakahori, Y., MacKinnon, R. N., Roche, A., Flint, T. J., Jacobs, P. A., Tommerup, N., Tranebjaerg, L., Froster-Iskenius, U., Kerr, B., Turner, B., Lindenbaum, R. H., Winter, R., Pembrey, M., Thibodeau, S. & Davies, K. E. (1991). **Physical mapping**

across the fragile X: hypermethylation and clinical expression of the fragile X syndrome. *Cell* **64**, 861-866.

Bennett, S. T., Lucassen, A. M., Gough, S. C., Powell, E. E., Undlien, D. E., Pritchard, L. E., Merriman, M. E., Kawaguchi, Y., Dronsfield, M. J., Pociot, F., Nerup, J., Bouzekri, N., Cambon-Thomsen, A., Rønningen, K. S., Bain, S. C. & Todd, J. A. (1995). **Susceptibility to human type 1 diabetes at *IDDM2* is determined by tandem repeat variation at the insulin gene minisatellite locus.** *Nature Genet.* **9**, 284-292.

Berger, I., Kang, C., Fredian, A., Ratliff, R., Moyzis, R. & Rich, A. (1995). **Extension of the four-stranded intercalated cytosine motif by adenine·adenine base pairing in the crystal structure of d(CCCAAT).** *Nature Struct. Biol.* **2**, 416-425.

Blommers, M. J. J., Haasnoot, C. A. G., Hilbers, C. W., van Boom, J. H. & van der Marel, G. A. (1987). **Structure and dynamics of biopolymers**. In *NATO ASI*, Series E: Applied Sciences No. 133, Martinus Nijhoff, Boston, pp. 78-91, quoted by Blommers *et al.*, 1989; Hilbers *et al.*, 1994.

Blommers, M. J. J., Walters, J. A. L. I., Haasnoot, C. A. G., Aelen, J. M. A., van der Marel, G. A., van Boom, J. H. & Hilbers, C. W. (1989). **Effects of base sequence on the loop folding in DNA hairpins.** *Biochemistry* **28**, 7,491-7,498.

Bouaziz, S., Kettani, A. & Patel, D. J. (1998). **A K cation-induced conformational switch within a loop spanning segment of a DNA quadruplex containing G-G-G-C repeats.** *J. Mol. Biol.* **282**, 637-652.

Bowater, R. P., Jaworski, A., Larson, J. E., Parniewski, P. & Wells, R. D. (1997). **Transcription increases the deletion frequency of long CTG center dot CAG triplet repeats from plasmids in Escherichia coli.** *Nucleic Acids Research* **25**, 2861-2868.

Brais, B., Bouchard, J.-P., Xie, Y.-G., Rochefort, D. L., Chrétien, N., Tomé, F. M., Lafrenière, R. G., Rommens, J. M., Uyama, E., Nohira, O., Blumen, S., Korcyn, A. D., Heutink, P., Mathieu, J., Duranceau, A., Codère, F., Fardeau, M. & Rouleau, G. A. (1998). **Short GCG expansions in the *PABP2* gene cause oculopharyngeal muscular dystrophy.** *Nature Genet.* **18**, 164-167.

Breschel, T. S., McInnis, M. G., Margolis, R. L., Sirugo, G., Corneliussen, B., Simpson, S. G., McMahon, F., MacKinnon, D. F., Xu, J. F., Pleasant, N., Huo, Y., Ashworth, R. G., Grundstrom, C., Grundstrom, T., Kidd, K. K., DePaulo, J. R. & Ross, C. A. (1997). **A novel, heritable, expanding CTG repeat in an intron of the SEF2-1 gene on chromosome 18q21.1.** *Hum. Mol. Genet.* **6**, 1,855-1,863.

Bronner, C. E., Baker, S. M., Morrison, P. T., Warren, G., Smith, L. G., Lescoe, M. K., Kane, M., Erabino, C., Lipford, J., Lindblom, A., Tannergård, P., Bollag, R. J., Godwin, A. R., Ward, D. C., Nordenskjøld, M., Fishel, R., Kolodner, R. & Liskay, R. M. (1994). **Mutation in the DNA mismatch repair gene homologue *hMLH1* is associated with hereditary non-polyposis colon cancer.** *Nature* **368**, 258-261.

Brook, J. D., McCurrach, M. E., Harley, H. G., Buckler, A. J., Church, D., Aburatani, H., Hunter, K., Stanton, V. P., Thirion, J. P., Hudson, T., Sohn, R., Zemelman, B., Snell, R. G., Rundle, S. A., Crow, S., Davies, J., Shelbourne, M., Buxton, J., Jones, C., Juvonen, V., Johnson, K., Harper, P. S., Shaw, D. J. & Housman, D. E. (1992). **Molecular basis of myotonic dystrophy: expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member.** *Cell* **68**, 799-808.

Brown, T. C., Tarleton, J. C., Go, R. C. P., Longshore, J. W. & Descartes, M. (1997). **Instability of the *FMR2* trinucleotide repeat region associated with expanded *FMR1* alleles.** *American Journal Of Medical Genetics* **73**, 447-455.

Buard, J., Bourdet, A., Yardley, J., Dubrova, Y. & Jeffreys, A. J. (1998). **Influences of array size and homogeneity on minisatellite mutation.** *EMBO J.* **17**, 3,495-3,502.

Buard, J. & Jeffreys, A. J. (1997). **Big, bad minisatellites.** *Nature Genet.* **15**, 327-328.

Buard, J. & Vergnaud, G. (1994). **Complex recombination events at the hypermutable minisatellite CEB1 (D2S90).** *EMBO J.* **13**, 3,203-3,210.

Burke, J. R., Wingfield, M. S., Lewis, K. E., Roses, A. D., Lee, J. E., Hulette, C., Pericak-Vance, M. A. & Vance, J. M. (1994). **The Haw River syndrome: dentatorubropallidoluysian atrophy (DRPLA) in an African-American family.** *Nature Genet.* **7**, 521-524.

Burright, E. N., Orr, H. T. & Clark, H. B. (1997). **Mouse models of human CAG repeat disorders.** *Brain Pathology* **7**, 965-977.

Buxton, J., Shelbourne, M., Davies, J., Jones, C., Van Tongeren, T., Aslanidis, C., de Jong, P., Jansen, G., Anvret, M., Riley, B., Williamson, R. & Johnson, K. (1992). **Detection of an unstable fragment of DNA specific to individuals with myotonic dystrophy.** *Nature* **355**, 547-548.

Campbell, A. (1965). **The stearic effect in lysogenization in bacteriophage lambda: 1. Lysogenization of a partially diploid strain of *Escherichia coli* K12.** *Virology* **27**, 329-339.

Campbell, T. A., Palmer, M. S., Will, R. G., Gibb, W. R., Luthert, P. J. & Collinge, J. (1996). **A prion disease with a novel 96-base pair insertional mutation in the prion protein gene.** *Neurology* **46**, 761-766.

Campuzano, V., Montermini, L., Moltò, M. D., Pianese, L., Cossée, M., Cavalcanti, F., Monros, E., Rodius, F., Duclos, F., Monticelli, A., Zara, F., Cañizares, J., Koutnikova, H., Bidichandani, S. I., Gellera, C., Brice, A., Truoillas, P., De Michele, G., Filla, A., De Frutos, R., Palau, F., Patel, P. I., Di Donato, S., Mandel, J.-L., Cocozza, S., Koenig, M. & Pandolfo, M. (1996). **Friedreich's ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion.** *Science* **271**, 1,423-1,427.

Cancel, G., Dürr, A., Didierjean, O., Imbert, G., Bürk, K., Lezin, A., Belal, S., Benomar, A., Abada-Bendib, M., Vial, C., Guimarães, J., Chneiweiss, H., Stevanin, G., Yvert, G., Abbas, N., Saudou, F., Lebre, A.-S., Yahyaoui, M., Hentati, F., Vernant, J.-C., Klockgether, T., Mandel, J.-L., Agid, Y. & Brice, A. (1997). **Molecular and clinical correlations in spinocerebellar ataxia 2: A study of 32 families.** *Hum. Mol. Genet.* **6**, 709-715.

Capellari, S., Vital, C., Parchi, P., Petersen, R. B., Ferrer, X., Jarnier, D., Pegoraro, E., Gambetti, P. & Julien, J. (1997). **Familial prion disease with a novel 144-bp insertion in the prion protein gene in a Basque family.** *Neurology* **49**, 133-141.

Capon, D. J., Chen, E. Y., D., L. A., Seeburg, P. H. & Goeddel, D. V. (1983). **Complete nucleotide sequences of the T24 human bladder carcinoma oncogene and its normal homologue.** *Nature* **302**, 33-37.

Carango, P., Noble, J. E., Marks, H. G. & Funanage, V. L. (1993). **Absence of myotonic dystrophy protein kinase (DMPK) mRNA as a result of a triplet repeat expansion in myotonic dystrophy.** *Genomics* **18**, 340-348.

Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J. & Deka, R. (1997). **Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci.** *Proc. Natl Acad. Sci. U.S.A.* **94**, 1,041-1,046.

Chalker, A. F., Leach, D. R. F. & Lloyd, R. G. (1988). ***Escherichia coli sbcC* mutants permit stable propagation of DNA replicons containing a long palindrome.** *Gene* **71**, 201-205.

Chalker, A. F., Okely, E. A., Davison, A. & Leach, D. R. F. (1993). **The effects of central asymmetry on the propagation of palindromic DNA in bacteriophage λ are consistent with cruciform extrusion *in vivo*.** *Genetics* **133**, 143-148.

Chamberlin, M. & Berg, P. (1962). **Deoxyribonucleic acid-directed synthesis of ribonucleic acid by an enzyme from *Escherichia coli*.** *Proc. Natl Acad. Sci., U.S.A.* **48**, 81-94.

Chang, C. S., Kokontis, J. & Liao, S. T. (1988). **Structural analysis of complementary DNA and amino acid sequences of human and rat androgen receptors.** *Proc. Natl Acad. Sci. U.S.A.* **85**, 7,211-7,215.

Chastain, P. D., II, Eichler, E. E., Kang, S., Nelson, D. L., Levene, S. D. & Sinden, R. R. (1995). **Anomalous rapid electrophoretic mobility of DNA containing triplet repeats associated with human disease genes.** *Biochemistry* **34**, 16,125-16,131.

Chastain, P. D. & Sinden, R. R. (1998). **CTG repeats associated with human genetic disease are inherently flexible.** *J. Mol. Biol.* **275**, 405-11.

Chen, F.-M. (1995). **Acid-facilitated supramolecular assembly of G-quadruplexes in $d(CGG)_4$.** *J. Biol. Chem.* **270**, 23,090-23,096.

Chen, X., Mariappan, S. V., Moyzis, R. K., Bradbury, E. M. & Gupta, G. (1998). **Hairpin induced slippage and hyper-methylation of the fragile X DNA triplets.** *J. Biomol. Struct. Dyn.* **15**, 745-756.

Chen, X., Mariappan, S. V. S., Catasti, P., Ratliff, R., Moyzis, R. K., Laayoun, A., Smith, S. S., Bradbury, E. M. & Gupta, G. (1995). **Hairpins are formed by the single DNA strands of the fragile X triplet repeats: Structure and biological implications.** *Proc. Natl Acad. Sci. U.S.A.* **92**, 5,199-5,203.

Cheng, X. & Blumenthal, R. M. (1996). **Finding a basis for flipping bases.** *Structure* **4**, 639-645.

Chung, M.-y., Ranum, L. P. W., Duvick, L. A., Servadio, A., Zoghbi, H. Y. & Orr, H. T. (1993). **Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I.** *Nature Genet.* **5**, 254-258.

Collinge, J. & Palmer, M. S. (1994). **Molecular genetics of human prion diseases.** *Philos. Trans. R. Soc. Lond. B* **343**, 371-378.

Cossée, M., Schmitt, M., Campuzano, V., Reutenauer, L., Moutou, C., Mandel, J.-L. & Koenig, M. (1997). **Evolution of the Friedreich's ataxia trinucleotide repeat expansion: founder effect and premutations.** *Proc. Natl Acad. Sci. U.S.A.* **94**, 7,452-7,457.

Darlow, J. M. & Leach, D. R. F. (1995). **The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in *Escherichia coli* suggest hairpin folding preferences *in vivo*.** *Genetics* **141**, 825-832.

Darlow, J. M. & Leach, D. R. F. (1998a). **Secondary structures in d(CGG) and d(CCG) repeat tracts.** *J. Mol. Biol.* **275**, 3-16.

Darlow, J. M. & Leach, D. R. F. (1998b). **Evidence for two preferred hairpin folding patterns in d(CGG).d(CCG) repeat tracts *in vivo*.** *J. Mol. Biol.* **275**, 17-23.

David, G., Abbas, N., Stevanin, G., Durr, A., Yvert, G., Cancel, G., Weber, C., Imbert, G., Saudou, F., Antoniou, E., Drabkin, H., Gemmill, R., Giunti, P., Benomar, A., Wood, N., Ruberg, M., Agid, Y., Mandel, J.-L. & Brice, A. (1997). **Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion.** *Nature Genet.* **17**, 65-70.

David, G., Dürr, A., Stevanin, G., Cancel, G., Abbas, N., Benomar, A., Belal, S., Lebre, A.-S., Abada-Bendib, M., Grid, D., Holmberg, M., Yahyaoui, M., Hentati, F., Chkili, T., Agid, Y. & Brice, A. (1998). **Molecular and clinical correlations in autosomal dominant cerebellar ataxia with progressive macular dystrophy (SCA7).** *Hum. Mol. Genet.* **7**, 165-170.

Davison, A. (1994). **DNA secondary structure *in vivo*.** Unpublished thesis, University of Edinburgh.

Davison, A. & Leach, D. R. F. (1994a). **The effects of nucleotide sequence changes on DNA secondary structure formation in *Escherichia coli* are consistent with cruciform extrusion *in vivo*.** *Genetics* **137**, 361-368.

Davison, A. & Leach, D. R. F. (1994b). **Two-base DNA hairpin-loop structures *in vivo*.** *Nucl. Acids Res.* **22**, 4,361-4,363.

Dawson, R. M. C., Elliott, D. C., Elliott, W. H. & Jones, K. M. (1986). *Data for biochemical research*. Third edit. Oxford Science Publications, Clarendon Press, Oxford.

de Boer, J. G. & Ripley, L. S. (1984). **Demonstration of the production of frameshift and base-substitution mutations by quasipalindromic DNA sequences.** *Proc. Natl Acad. Sci., U.S.A.* **81**, 5,528-5,531.

Del-Favero, J., Krols, L., Michalik, A., Theuns, J., Löfgren, A., Goossens, D., Wehnert, A., Van den Bossche, D., Van Zand, K., Backhovens, H., van Regenmorter, N., Martin, J.-J. & Van Broeckhoven, C. (1998). **Molecular genetic analysis of autosomal dominant cerebellar ataxia with retinal degeneration (ADCA type II) caused by CAG triplet repeat expansion.** *Hum. Mol. Genet.* **7**, 177-186.

DePamphilis, M. L. & Wassarman, P. M. (1980). **Replication of eukaryotic chromosomes: a close-up of the replication fork.** *Annu. Rev. Biochem.* **49**, 627-666.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M. & Freimer, N. B. (1994). **Mutational processes of simple-sequence repeat loci in human populations.** *Proc. Natl Acad. Sci. U.S.A.* **91**, 3,166-3,170.

Dietrich, A., Kioschis, P., Monaco, A. P., Gross, B., Korn, B., Williams, S. V., Sheer, D., Heitz, D., Oberlé, I., Toniolo, D., Warren, S. T., Lehrach, H. & Poustka, A. (1991). **Molecular cloning and analysis of the fragile X region in Man.** *Nucl. Acids Res.* **19**, 2,567-2,572.

Doyu, M., Sobue, G., Mukai, E., Kachi, T., Yasuda, T., Mitsuma, T. & Takahashi, A. (1992). **Severity of X-linked recessive bulbospinal neuronopathy correlates with size of the tandem CAG repeat in androgen receptor gene.** *Ann. Neurol.* **32**, 707-710.

Dürr, A., Stevanin, G., Cancel, G., Duyckaerts, C., Abbas, N., Didierjean, O., Chneiweiss, H., Benomar, A., Lyon-Caen, O., Julien, J., Serdaru, M., Penet, C., Agid, Y. & Brice, A. (1996). **Spinocerebellar ataxia 3 and Machado-Joseph disease: clinical, molecular, and neuropathological features.** *Ann. Neurol.* **39**, 490-499.

Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T. & Chakraborty, R. (1992). **Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups.** *Genomics* **12**, 241-253.

Eichler, E. E., Holden, J. J. A., Popovich, B. W., Reiss, A. L., Snow, K., Thibodeau, S. N., Richards, C. S., Ward, P. A. & Nelson, D. L. (1994). **Length of uninterrupted CGG repeats determines instability in the *FMR1* gene.** *Nature Genet.* **8**, 88-94.

Eichler, E. E., Macpherson, J. N., Murray, A., Jacobs, P. A., Chakravarti, A. & Nelson, D. L. (1996). **Haplotype and interspersion analysis of the FMR1 CGG repeat identifies two different mutational pathways for the origin of the fragile X syndrome.** *Hum. Mol. Genet.* **5**, 319-330.

Epplen, C. & Epplen, J. T. (1994). **Expression of $(cac)_n/(gtg)_n$ simple repetitive sequences in mRNA of human lymphocytes.** *Hum. Genet.* **93**, 35-41.

Epplen, J. T., Ammer, H., Kammerbauer, C., Schwaiger, W., Schmid, M. & Nanda, I. (1991). **On the meaning of hypervariable simple repetitive DNA loci in eukaryotic genomes: an initial attempt for a basic theoretical assesment.** *Advances in Mol. Gen.* **4**, 301-310.

Falaschi, A., Adler, J. & Khorana, H. G. (1963). **Chemically synthesized deoxypolynucleotides as templates for ribonucleic acid polymerase.** *J. Biol. Chem.* **238**, 3,080-3,085.

Farabaugh, P. J., Schmeissner, U., Hofer, M., Miller, J. H. & by, G. D. J. w. a. (1978). **Genetic studies of the *lac* repressor VII. On the molecular nature of spontaneous hotspots in the *lacI* gene of *Escherichia coli*.** *J. Mol. Biol.* **126**, 847-863.

Fearon, E. R., Cho, K. R., Nigro, J. M., Kern, S. E., Simons, J. W., Ruppert, J. M., Hamilton, S. R., Preisinger, A. C., Thomas, G., Kinzler, K. W. & et al. (1990). **Identification of a chromosome 18q gene that is altered in colorectal cancers.** *Science* **247**, 49-56.

Fisch, G. S., Snow, K., Thibodeau, S. N., Chalifaux, M., Holden, J. J. A., Nelson, D. L., Howard-Peebles, P. N. & Maddalena, A. (1995). **The fragile X premutation in carriers and its effect on mutation size in offspring.** *Am. J. Hum. Genet.* **56**, 1,147-1,155.

Fishel, R., Ewel, A., Lee, S., Lescoe, M. K. & Griffith, J. (1994). **Binding of mismatched microsatellite DNA sequences by the human MSH2 protein.** *Science* **266**, 1,403-1,405.

Fishel, R., Lescoe, M. K., Rao, M. R. S., Copeland, N. G., Jenkins, N. A., Garber, J., Kane, M. & Kolodner, R. (1993). **The human mutator gene homologue *MSH2* and its association with hereditary nonpolyposis colon cancer.** *Cell* **75**, 1,027-1,038.

Fraser, F. C. (1997). **Trinucleotide repeats not the only cause of anticipation.** *Lancet* **350**, 459-460.

Fresco, J. R. & Alberts, B. M. (1960). **The accommodation of non-complementary bases in helical polyribonucleotides and deoxyribonucleic acids.** *Proc. Natl Acad. Sci. U.S.A.* **46**, 311-321.

Freudenreich, C. H., Stavenhagen, J. B. & Zakian, V. A. (1997). **Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome.** *Mol. Cell. Biol.* **17**, 2,090-2,098.

Frontali, C., Dore, E., Feranto, A., Gratton, E., Bettini, A., Pozzan, M. R. & Valdevit, E. (1979). **An absolute method for the determination of the persistence length of native DNA from electron micrographs.** *Biopolymers* **18**, 1,353-1,373.

Fry, M. & Loeb, L. A. (1994). **The fragile X syndrome $d(CGG)_n$ nucleotide repeats form a stable tetrahelical structure.** *Proc. Natl Acad. Sci. U.S.A.* **91**, 4,950-4,954.

Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Pieretti, M., Sutcliffe, J. S., Richards, S., Verkerk, A. J. M. H., Holden, J. J. A., Fenwick, R. G. J., Warren, S. T., Oostra, B. A., Nelson, D. L. & Caskey, C. T. (1991). **Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox.** *Cell* **67**, 1,047-1,058.

Fu, Y.-H., Pizzuti, A., Fenwick, R. G., Jr., King, J., Rajnarayan, S., Dunne, P. W., Dubel, J., Nasser, G. A., Ashizawa, T., de Jong, P., Weiringa, B., Korneluk, R., Perryman, M. B., Epstein, H. F. & Caskey, C. T. (1992). **An unstable triplet repeat in a gene related to myotonic muscular dystrophy.** *Science* **255**, 1,256-1,258.

Gacy, A. M., Goellner, G., Juranic, N., Macura, S. & McMurray, C. T. (1995). **Trinucleotide repeats that expand in human disease form hairpin structures in vitro.** *Cell* **81**, 533-540.

Gacy, A. M., Goellner, G. M., Spiro, C., Chen, X., Gupta, G., Bradbury, E. M., Dyer, R. B., Mikesell, M. J., Yao, J. Z., Johnson, A. J., Richter, A., Melançon, S. B. & McMurray, C. T. (1998). **GAA instability in Friedreich's Ataxia shares a common, DNA-directed and intraallelic mechanism with other trinucleotide diseases.** *Molecular Cell* **1**, 583-593.

Gacy, A. M. & McMurray, C. T. (1998). **Influence of hairpins on template reannealing at trinucleotide repeat duplexes: a model for slipped DNA.** *Biochemistry* **37**, 9,426-9,434.

Gallego, J., Chou, S.-H. & Reid, B. R. (1997). **Centromeric pyrimidine strands fold into an intercalated motif by forming a double hairpin with a novel T:G:G:T tetrad: solution structure of the d(TCCCGTTTCCA) dimer.** *J. Mol. Biol.* **273**, 840-856.

Gao, X., Huang, X., Smith, G. K., Zheng, M. & Liu, H. (1995). **New antiparallel duplex motif of DNA CCG repeats that is stabilised by extrahelical bases symmetrically located in the minor groove.** *J. Am. Chem. Soc.* **117**, 8,883-8,884.

Gecz, J., Gedeon, A. K., Sutherland, G. R. & Mulley, J. C. (1996). **Identification of the gene *FMR2*, associated with FRAXE mental retardation.** *Nature Genet.* **13**, 105-108.

Gehring, K., Leroy, J.-L. & Guéron, M. (1993). **A tetrameric DNA structure with protonated cytosine·cytosine base pairs.** *Nature* **363**, 561-565.

Gellibolian, R., Bacolla, A. & Wells, R. D. (1997). **Triplet repeat instability and DNA topology: an expansion model based on statistical mechanics.** *J. Biol. Chem.* **272**, 16,793-16,797.

Gervais, V., Cognet, J. A. H., Le Bret, M., Sowers, L. C. & Fazakerley, G. V. (1995). **Solution structure of two mismatches A·A and T·T in the *K-ras* gene context by nuclear magnetic resonance and molecular dynamics.** *Eur. J. Biochem.* **228**, 279-290.

Gibbs, M., Collick, A., Kelly, R. G. & Jeffreys, A. J. (1993). **A tetranucleotide repeat mouse minisatellite displaying substantial somatic instability during early preimplantation development.** *Genomics* **17**, 121-128.

Gibson, F. P., Leach, D. R. F. & Lloyd, R. G. (1992). **Identification of *sbcD* mutations as cosuppressors of *recBC* that allow propagation of DNA palindromes in *Escherichia coli* K-12.** *J. Bacteriol.* **174**, 1,222-1,228.

Goellner, G. M., Tester, D., Thibodeau, S., Almqvist, E., Goldberg, Y. P., Hayden, M. R. & McMurray, C. T. (1997). **Different mechanisms underlie DNA instability in Huntington disease and colorectal cancer.** *Am. J. Hum. Genet.* **60**, 879-890.

Goldfarb, L. G., Brown, P., Cervenakova, L. & Gajdusek, D. C. (1994). **Genetic analysis of Creutzfeldt-Jakob disease and related disorders.** *Philos. Trans. R. Soc. Lond. B* **343**, 379-384.

Goodman, F. R., Mundlos, S., Muragaki, Y., Donnai, D., Giovannucci-Uzielli, M. L., Lapi, E., Majewski, F., McGaughran, J., McKeown, C., Reardon, W., Upton, J., Winter, R. M., Olsen, B. R. & Scambler, P. J. (1997). **Synpolydactyly phenotypes correlate with size of expansions in HOXD13 polyalanine tract.** *Proc. Natl Acad. Sci. U.S.A.* **94**, 7,458-7,463.

Gordenin, D. A., Kunkel, T. A. & Resnick, M. A. (1997). **Repeat expansion--all in a flap?** *Nature Genet.* **16**, 116-118.

Gouw, L. G., Castañeda, M. A., McKenna, C. K., Digre, K. B., Pulst, S. M., Perlman, S., Lee, M. S., Gomez, C., Fischbeck, K., Gagnon, D., Storey, E., Bird, T., Jeri, F. R. & Ptácek, L. J. (1998). **Analysis of the dynamic mutation in the SCA7 gene shows marked parental effects on CAG repeat transmission.** *Hum. Mol. Genet.* **7**, 525-532.

Green, H. & Wang, N. (1994). **Codon reiteration and the evolution of proteins.** *Proc. Natl Acad. Sci. U.S.A.* **91**, 4,298-4,302.

Green, M. & Krontiris, T. G. (1993). **Allelic variation of reporter gene activation by the HRAS1 minisatellite.** *Genomics* **17**, 429-434.

Gu, Y., Shen, Y., Gibbs, R. A. & Nelson, D. L. (1996). **Identification of *FMR2*, a novel gene associated with the *FRAXE* CCG repeat and CpG island.** *Nature Genet.* **13**, 109-113.

Hammond-Kosack, M. C. U., Dobrinski, B., Lurz, R., Docherty, K. & Kilpatrick, M. W. (1992a). **The human insulin gene linked polymorphic region exhibits an altered DNA structure.** *Nucl. Acids Res.* **20**, 231-236.

Hammond-Kosack, M. C. U. & Docherty, K. (1992b). **A consensus repeat sequence from the human insulin gene linked polymorphic region adopts multiple quadriplex DNA structures *in vitro*.** *FEBS Lett.* **301**, 79-82.

Hammond-Kosack, M. C. U., Kilpatrick, M. W. & Docherty, K. (1992c). **Analysis of DNA structure in the human insulin gene-linked polymorphic region *in vivo*.** *J. Mol. Endocrinol.* **9**, 221-225.

Hammond-Kosack, M. C. U., Kilpatrick, M. W. & Docherty, K. (1993). **The human insulin gene-linked polymorphic region adopts a G-quartet structure in chromatin assembled *in vitro*.** *J. Mol. Endocrinol.* **10**, 121-126.

Hansen, R. S., Canfield, T. K., Lamb, M. M., Gartler, S. M. & Laird, C. D. (1993). **Association of fragile-X syndrome with delayed replication of the *FMR1* gene.** *Cell* **73**, 1,403-1,409.

Hanvey, J. C., Shimizu, M. & Wells, R. D. (1989). **Intramolecular DNA triplexes in supercoiled plasmids. II. Effect of base composition and noncentral interruptions on formation and stability.** *J. Biol. Chem.* **264**, 5,950-5,956.

Hardin, C. C., Corregan, M., Brown, B. A. I. & Frederick, L. N. (1993). **Cytosine-cytosine$^+$ base pairing stabilizes DNA quadruplexes and cytosine methylation greatly enhances the effect.** *Biochemistry* **32**, 5,870-5,880.

Hardin, C. C., Henderson, E., Watson, T. & Prosser, J. K. (1991). **Monovalent cation induced structural transitions in telomeric DNAs: G-DNA folding intermediates.** *Biochemistry* **30**, 4,460-4,472.

Hardin, C. C., Watson, T., Corregan, M. & Bailey, C. (1992). **Cation-dependent transition between the quadruplex and Watson-Crick hairpin forms of d(CGCG$_3$GCG).** *Biochemistry* **31**, 833-841.

Harley, H. G., Brook, J. D., Rundle, S. A., Crow, S., Reardon, W., Buckler, A. J., Harper, P. S., Housman, D. E. & Shaw, D. J. (1992a). **Expansion of an unstable DNA region and phenotypic variation in myotonic dystrophy.** *Nature* **355**, 545-546.

Harley, H. G., Rundle, S. A., Reardon, W., Myring, J., Crow, S., Brook, J. D., Harper, P. S. & Shaw, D. J. (1992b). **Unstable DNA sequence in myotonic dystrophy.** *Lancet* **339**, 1,125-1,128.

Harper, P. S. (1989). *Myotonic dystrophy*, W.B. Saunders Co., London.

Harper, P. S. (1997). **Trinucleotide repeat disorders.** *J. Inher. Metab. Dis.* **20**, 122-124.

Harper, P. S., Harley, H. G., Reardon, W. & Shaw, D. J. (1992). **Anticipation in myotonic dystrophy: new light on an old problem.** *Am. J. Hum. Genet.* **51**, 10-16.

Harris, S., Moncrieff, C. & Johnson, K. (1996). **Myotonic dystrophy: will the real gene please step forward!** *Hum. Mol. Genet.* **5**, 1,417-1,423.

Harvey, S. C. (1997). **Slipped structures in DNA triplet repeat sequences: entropic contributions to genetic instabilities.** *Biochemistry* **36**, 3,047-3,049.

Hastings, P. J. (1988). **Recombination in the eukaryotic nucleus.** *BioEssays* **9**, 61-64.

Heale, S. M. & Petes, T. D. (1995). **The stabilization of repetitive tracts of DNA by variant repeats requires a functional DNA mismatch repair system.** *Cell* **83**, 539-545.

Heitz, D., Rousseau, F., Devys, D., Saccone, S., Abderrahim, H., Le Paslier, D., Cohen, D., Vincent, A., Toniolo, D., Della Valle, G., Johnson, S., Schlessinger, D., Oberlé, I. & Mandel, J.-L. (1991). **Isolation of sequences that span the fragile-X and identification of a fragile-X related CpG island.** *Science* **251**, 1,236-1,239.

Henderson, E. R., Moore, M. & Malcolm, B. A. (1990). **Telomere G-strand structure and function analyzed by chemical protection, base analogue substitution, and utilization by telomerase *in vitro*.** *Biochemistry* **29**, 732-737.

Hentschel, C. C. (1982). **Homocopolymer sequences in the spacer of a sea urchin histone gene repeat are sensitive to S_1 nuclease.** *Nature* **295**, 714-716.

Hewett, D. R., Handt, O., Hobson, L., Mangelsdorf, M., Eyre, H. J., Baker, E., Sutherland, G. R., Schuffenhauer, S., Mao, J.-I. & Richards, R. I. (1998). ***FRA10B* structure reveals common elements in repeat expansion and chromosomal fragile site genesis.** *Mol. Cell* **1**, 773-781.

Hilbers, C. W., Heus, H. A., van Dongen, M. J. P. & Wijmenga, S. S. (1994). **The hairpin elements of nucleic acid structure: DNA and RNA folding.** *Nucleic Acids and Mol. Biol.* **8**, 56-104.

Hirst, M. C., Grewal, P. K. & Davies, K. E. (1994). **Precursor arrays for triplet repeat expansion at the fragile X locus.** *Hum, Mol. Genet.* **3**, 1,553-1,560.

Hohn, B. (1979). ***In vitro* packaging of lambda and cosmid DNA.** *Methods Enzymol.* **68**, 299-309.

Holden, J. J. A., Walker, M., Chalifoux, M. & White, B. N. (1996). **Trinucleotide repeats at the FRAXF locus: frequency and distribution in the general population.** *Am. J. Med. Genet.* **64**, 424-427.

Horwitz, M., Goode, E. L. & Jarvik, G. P. (1996). **Anticipation in familial leukemia.** *Am. J. Hum. Genet.* **59**, 990-998.

Howard-Flanders, P. & Theriot, l. (1966). **Mutants of *Escherichia coli* K-12 defective in DNA repair and genetic recombination.** *Genetics* **53**, 1137-1150.

Höweler, C. J., Busch, H. F., Geraedts, J. P., Niermeijer, M. F. & Staal, A. (1989). **Anticipation in myotonic dystrophy: fact or fiction?** *Brain* **112**, 779-797.

Howell, R. M., Woodford, K. J., Weitzmann, M. N. & Usdin, K. (1996). **The chicken β-globin gene promoter forms a novel "cinched" tetrahelical structure.** *J. Biol. Chem.* **271**, 5,208-5,214.

Hummerich, H., Baxendale, S., Mott, R., Kirby, S. F., MacDonald, M. E., Gusella, J., Lehrach, H. & Bates, G. P. (1994). **Distribution of trinucleotide repeat sequences across a 2 Mbp region containing the Huntington's disease gene.** *Hum. Mol. Genet.* **3**, 73-78.

Igarashi, S., Takiyama, Y., Cancel, G., Rogaeva, E. A., Sasaki, H., Wakisaka, A., Zhou, Y.-X., Takano, H., Endo, K., Sanpei, K., Oyake, M., Tanaka, H., Stevanin, G., Abbas, N., Dürr, A., Rogaev, E. I., Sherrington, R., Tsuda, T., Ikeda, M., Cassa, E., Nishizawa, M., Benomar, A., Julien, J., Weissenbach, J., Wang, G.-X., Agid, Y., St. George-Hyslop, P. H., Brice, A. & Tsuji, S. (1996). **Intergenerational instability of the CAG repeat of the gene for Machado- Joseph disease (*MJD1*) is affected by the genotype of the normal chromosome: implications for the molecular mechanisms of the instability of the CAG repeat.** *Hum. Mol. Genet.* **5**, 923-932.

Igarashi, S., Tanno, Y., Onodera, O., Yamazaki, M., Sato, S., Ishikawa, A., Miyatani, N., Nagashima, M., Ishikawa, Y., Sahashi, K., Ibi, T., Miyatake, T. & Tsuji, S. (1992). **Strong correlation between the number of CAG repeats in androgen receptor genes and the clinical onset of features of spinal and bulbar muscular atrophy.** *Neurology* **42**, 2,300-2,302.

Ikeuchi, T., Takano, H., Koide, R., Horikawa, Y., Honma, Y., Onishi, Y., Igarashi, S., Tanaka, H., Nakao, N., Sahashi, K., Tsukagoshi, H., Inoue, K., Takahashi, H. & Tsuji, S. (1997). **Spinocerebellar ataxia type 6: CAG repeat expansion in α_{1A} voltage-dependent calcium channel gene and clinical variations in Japanese population.** *Ann. Neurol.* **42**, 879-884.

Imbert, G., Kretz, C., Johnson, K. & Mandel, J.-L. (1993). **Origin of the expansion mutation in myotonic dystrophy.** *Nature Genet.* **4**, 72-76.

Imbert, G., Saudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J.-M., Weber, C., Mandel, J.-L., Cancel, G., Abbas, N., Dürr, A., Didierjean, O., Stevanin, G., Agid, Y. & Brice, A. (1996). **Cloning of the gene For spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats.** *Nature Genet.* **14**, 285-291.

Ishikawa, K., Tanaka, H., Saito, M., Ohkoshi, N., Fujita, T., Yoshizawa, K., Ikeuchi, T., Watanabe, M., Hayashi, A., Takiyama, Y., Nishizawa, M., Nakano, I., Matsubayashi, K., Miwa, M., Shoji, S., Kanazawa, I., Tsuji, S. & Mizusawa, H. (1997). **Japanese families with autosomal dominant pure cerebellar ataxia map to chromosome 19p13.1-p13.2 and are strongly associated with mild CAG expansions in the spinocerebellar ataxia type 6 gene in chromosome 19p13.1.** *Am. J. Hum. Genet.* **61**, 336-346.

Ivonov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. (1993). **Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis.** *Nature* **363**, 558-561.

Jansen, G., Willems, P., Coerwinkel, M., Nillesen, W., Smeets, H., Vits, L., Höweler, C., Brunner, H. & Weiringa, B. (1994). **Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic events in $(CTG)_n$ repeat variations and selection against extreme expansion in sperm.** *Am. J. Hum. Genet.* **54**, 575-585.

Jaworski, A., Rosche, W. A., Gellibolian, R., Kang, S., Shimizu, M., Bowater, R. P., Sinden, R. R. & Wells, R. D. (1995). **Mismatch repair in *Escherichia coli* enhances instability of $(CTG)_n$ triplet repeats from human hereditary diseases.** *Proc. Natl Acad. Sci. U.S.A.* **92**, 11019-11023.

Jeffreys, A. J., MacLeod, A., Tamaki, K., Neil, D. L. & Monckton, D. G. (1991). **Minisatellite repeat coding as a digital approach to DNA typing.** *Nature* **204**, 204-209.

Jeffreys, A. J., Neil, D. L. & Neumann, R. (1998). **Repeat instability at human minisatellites arising from meiotic recombination.** *EMBO J.* **17**, 4,147-4,157.

Jeffreys, A. J. & Neumann, R. (1997). **Somatic mutation processes at a human minisatellite.** *Hum Mol Genet* **6**, 129-32; 134-6.

Jeffreys, A. J., Neumann, R. & Wilson, V. (1990). **Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis.** *Cell* **60**, 473-485.

Jeffreys, A. J., Royle, N. J., Wilson, V. & Wong, Z. (1988). **Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA.** *Nature* **332**, 278-281.

Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, D. G., Neil, D. L. & Armour, J. A. L. (1994). **Complex gene conversion events in germline mutation at human minisatellites.** *Nature Genet.* **6**, 136-145.

Jeffreys, A. J., Wilson, V., Kelly, R., Taylor, B. A. & Bulfield, G. (1987). **Mouse DNA 'fingerprints': analysis of chromosome localization and germ-line stability of hypervariable loci in recombinant inbred strains.** *Nucl. Acids Res.* **15**, 2,823-2,836.

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985). **Hypervariable 'minisatellite' regions in human DNA.** *Nature* **314**, 67-73.

Johansson, J., Forsgren, L., Sandgren, O., Brice, A., Holmgren, G. & Holmberg, M. (1998). **Expanded CAG repeats in Swedish spinocerebellar ataxia type 7 (SCA7) patients: effect of CAG repeat length on the clinical manifestation.** *Hum. Mol. Genet.* **7**, 171-176.

Jollès, B., Réfrégiers, M. & Laigle, A. (1997). **Opening of the extraordinarily stable mini-hairpin d(GCGAAGC).** *Nucl. Acids Res.* **25**, 4,608-4,613.

Jones, C., Penny, L., Mattina, T., Yu, S., Baker, E., Voullaire, L., Langdon, W. Y., Sutherland, G. R., Richards, R. I. & Tunnacliffe, A. (1995). **Association of a chromosome deletion syndrome with a fragile site within the proto-oncogene *CBL2*.** *Nature* **376**, 145-149.

Jones, C., Slijepcevic, P., Marsh, S., Baker, E., Langdon, W. Y., Richards, R. I. & Tunnacliffe, A. (1994). **Physical linkage of the fragile site *FRA11B* and a Jacobsen syndrome chromosome deletion breakpoint in 11q23.3.** *Hum. Mol. Genet.* **3**, 2,123-2,130.

Kakizuka, A. (1997). **Degenerative ataxias: genetics, pathogenesis and animal models.** *Curr. Opin. Neurol.* **10**, 285-290.

Kang, S., Jaworski, A., Ohshima, K. & Wells, R. D. (1995). **Expansion and deletion of CTG repeats from human disease genes are determined by the direction of replication in *E. coli*.** *Nature Genet.* **10**, 213-218.

Kang, S., Ohshima, K., Jaworski, A. & Wells, R. D. (1996). **CTG triplet repeats from the myotonic dystrophy gene are expanded in *Escherichia coli* distal to the replication origin as a single large event.** *J. Mol. Biol.* **258**, 543-547.

Kang, S., Ohshima, K., Shimizu, M., Amirhaeri, S. & Wells, R. D. (1995). **Pausing of DNA synthesis *in vitro* at specific loci in CTG and CGG triplet repeats from human hereditary disease genes.** *J. Biol. Chem.* **270**, 27,014-27,021.

Kawaguchi, Y., Okamoto, T., Taniwaki, M., Aizawa, M., Inoue, M., Katayama, S., Kawakami, H., Nakamura, S., Nishimura, M., Akiguchi, I., Kimura, J., Narumiya, S. & Kakizuka, A. (1994). **CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1.** *Nature Genet.* **8**, 221-228.

Kawakami, H., Maruyama, H., Nakamura, S., Kawaguchi, Y., Kakizuka, A., Doyu, M. & Sobue, G. (1995). **Unique features of the CAG repeats in Machado-Joseph disease.** *Nature Genet.* **9**, 344-345.

Kelly, R., Bulfield, G., Collick, A., Gibbs, M. & Jeffreys, A. J. (1989). **Characterization of a highly unstable mouse minisatellite locus: Evidence for somatic mutation during early development.** *Genomics* **5**, 844-856.

Kennedy, G. C., German, M. S. & Rutter, W. J. (1995). **The minisatellite in the diabetes susceptibility locus *IDDM2* regulates insulin transcription.** *Nature Genet.* **9**, 293-298.

Kettani, A., Bouaziz, S., Gorin, A., Zhao, H., Jones, R. A. & Patel, D. J. (1998). **Solution structure of a Na cation stabilized DNA quadruplex containing G·G·G·G and G·C·G·C tetrads formed by G-G-G-C repeats observed in adeno-associated viral DNA.** *J. Mol. Biol.* **282**, 619-636.

Kettani, A., Bouaziz, S., Wang, W., Jones, R. A. & Patel, D. J. (1997). ***Bombyx mori* single repeat telomeric DNA sequence forms a G-quadruplex capped by base triads.** *Nature Struct. Biol.* **4**, 382-389.

Kettani, A., Kumar, R. A. & Patel, D., J. (1995). **Solution structure of a DNA quadruplex containing the fragile X syndrome triplet repeat.** *J. Mol. Biol.* **254**, 638-656.

Kho, M. R., Baker, D. J., Laayoun, A. & Smith, S. S. (1998). **Stalling of human DNA (cytosine-5) methyltransferase at single-strand conformers from a site of dynamic mutation.** *J. Mol. Biol.* **275**, 67-79.

Khorana, H. G., Büchi, H., Ghosh, H., Gupta, N., Jacob, T. M., Kössel, H., Morgan, R., Narang, S. A., Ohtsuka, E. & Wells, R. D. (1966). **Polynucleotide synthesis and the genetic code.** *Cold Spring Harb. Symp. Quant. Biol.* **31**, 39-49.

Klockgether, T. & Dichgans, J. (1997). **Trinucleotide repeats and hereditary ataxias.** *Nature Med.* **3**, 149-150.

Knight, S. J. L., Flannery, A. V., Hirst, M. C., Campbell, L., Christodoulou, Z., Phelps, S. R., Pointon, J., Middleton-Price, S. R., Barnicoat, A., Pembrey, M. E., Holland, J., Oostra, B. A., Bobrow, M. & Davies, K. E. (1993). **Trinucleotide repeat amplification and hypermethylation of a CpG island in *FRAXE* mental retardation.** *Cell* **74**, 127-134.

Koeppen, A. H. (1998). **The hereditary ataxias.** *J. Neuropathol. Exp. Neurol.* **57**, 531-543.

Kohwi, Y., Wang, H. & Kohwi-Shigematsu, T. (1993). **A single trinucleotide, 5'AGC3'/5'GCT3', of the triplet-repeat disease genes confers metal ion-induced non-B DNA structure.** *Nucl. Acids Res.* **21**, 5,651-5,655.

Koide, R., Ikeuchi, T., Onodera, O., Tanaka, H., Igarashi, S., Endo, K., Takahashi, H., Kondo, R., Ishikawa, A., Hayashi, T., Saito, M., Tomoda, A., Miike, T., Naito, H., Ikuta, F. & Tsuji, S. (1994). **Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA).** *Nature Genet.* **6**, 9-13.

Kornberg, A., Bertsch, L. L., Jackson, J. F. & Khorana, H. G. (1964). **Enzymatic synthesis of deoxyribonucleic acid, XVI. oligonucleotides as templates and the mechanism of their replication.** *Proc. Natl Acad. Sci., U.S.A.* **51**, 315-323.

Korneluk, R. G. & Narang, M. A. (1997). **Anticipating anticipation.** *Nature Genet.* **15**, 119-120.

Koshy, B. T. & Zoghbi, H. Y. (1997). **The CAG/polyglutamine tract diseases: gene products and molecular pathogenesis.** *Brain Pathol.* **7**, 927-942.

Krahe, R., Eckhart, M., Ogunniyi, A. O., Osuntokun, B. O., Siciliano, M. J. & Ashizawa, T. (1995). **De novo myotonic dystrophy mutation in a Nigerian kindred.** *Am. J. Hum. Genet.* **56**, 1,067-1,074.

Kramer, P. R., Pearson, C. E. & Sinden, R. R. (1996). **Stability of triplet repeats of myotonic dystrophy and fragile X loci in human mutator mismatch repair cell lines.** *Hum. Genet.* **98**, 151-157.

Krasemann, S., Zerr, I., Weber, T., Poser, S., Kretzschmar, H., Hunsmann, G. & Bodemer, W. (1995). **Prion disease associated with a novel nine octapeptide repeat insertion in the PRNP gene.** *Brain Res. Mol. Brain. Res.* **34**, 173-176.

Kremer, E. J., Pritchard, M., Lynch, M., Yu, S., Holman, K., Baker, E., Warren, S. T., Schlessinger, D., Sutherland, G. R. & Richards, R. I. (1991). **Mapping of DNA instability at the fragile X to a trinucleotide repeat sequence p(CCG)$_n$.** *Science* **252**, 1,711-1,714.

Krontiris, T. G., Devlin, B., Karp, D. D., Robert, N. J. & Risch, N. (1993). **An association between the risk of cancer and mutations in the HRAS1 minisatellite locus.** *N. Engl. J. Med.* **329**, 517-523.

Kubitschek, H. E. & Henderson, T. R. (1966). **DNA replication.** *Proc. Natl Acad. Sci. U.S.A.* **55**, 512-519.

Kuhl, D. P. A. & Caskey, C. T. (1993). **Trinucleotide repeats and genome variation.** *Curr. Opin. Genet. Dev.* **3**, 404-407.

Kunkel, T. A. (1993). **Slippery DNA and diseases.** *Nature* **365**, 207-208.

Kunst, C. B. & Warren, S. T. (1994). **Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles.** *Cell* **77**, 853-861.

Kurohara, K., Kuroda, Y., Maruyama, H., Kawakami, H., Yukitake, M., Matsui, M. & Nakamura, S. (1997). **Homozygosity for an allele carrying intermediate CAG repeats in the dentatorubral-pallidoluysian atrophy (DRPLA) gene results in spastic paraplegia.** *Neurology* **48**, 1,087-1,090.

Kuryavyi, V. V. & Jovin, T. M. (1995a). **Triad-DNA.** *J. Biomol. Struct. Dyn.* **12**, a126.

Kuryavyi, V. V. & Jovin, T. M. (1995b). **Triad-DNA: a model for trinucleotide repeats.** *Nature Genet.* **9**, 339-341.

La Spada, A. R. (1997). **Trinucleotide repeat instability: genetic features and molecular mechanisms.** *Brain Pathol.* **7**, 943-963.

La Spada, A. R., Roling, D. B., Harding, A. E., Warner, C. L., Spiegel, R., Hausmanowa-Petrusewicz, I., Yee, W. C. & Fischbeck, K. H. (1992). **Meiotic stability and genotype-phenotype correlation of the trinucleotide repeat in X-linked spinal and bulbar muscular atrophy.** *Nature Genet.* **2**, 301-304.

La Spada, A. R., Wilson, E. M., Lubahn, D. B., Harding, A. E. & Fischbeck, K. H. (1991). **Androgen receptor gene mutations in X-linked spinal and bulbar muscular atrophy.** *Nature* **352**, 77-79.

Laayoun, A. & Smith, S. S. (1995). **Methylation of slipped duplexes, snapbacks and cruciforms by human DNA(cytosine-5)methyltransferase.** *Nucl. Acids Res.* **23**, 1,584-1,589.

Lafrenière, R. G., Rochefort, D. L., Chrétien, N., Rommens, J. M., Cochius, J. I., Kälviäinen, R., Nousiainen, U., Patry, G., Farrell, K., Söderfeldt, B., Federico, A., Hale, B. R., Cossio, O. H., Sørensen, T., Pouliot, M. A., Kmiec, T., Uldall, P., Janszky, J., Pranzatelli, M. R., Andermann, F., Andermann, E. & Rouleau, G. A. (1997). **Unstable insertion in the 5' flanking region of the cystatin B gene is the most common mutation in progressive myoclonus epilepsy type 1, EPM1.** *Nature Genet.* **15**, 298-302.

Laird, C. D. (1987). **Proposed mechanism of inheritance and expression of the human fragile-X syndrome of mental retardation.** *Genetics* **117**, 587-599.

Lalioti, M. D., Mirotsou, M., Buresi, C., Peitsch, M. C., Rossier, C., Ouazzani, R., Baldy-Moulinier, M., Bottani, A., Malafosse, A. & Antonarakis, S. E. (1997a). **Identification of mutations in *cystatin B*, the gene responsible for the Unverricht-Lundborg type of progressive myoclonus epilepsy (EPM1).** *Am. J. Hum. Genet.* **60**, 342-351.

Lalioti, M. D., Scott, H. S. & Antonarakis, S. E. (1997c). **What is expanded in progressive myoclonus epilepsy?** *Nature Genet.* **17**, 17.

Lalioti, M. D., Scott, H. S., Buresi, C., Rossier, C., Bottani, A., Morris, M. A., Malafosse, A. & Antonarakis, S. E. (1997b). **Dodecamer repeat expansion in cystatin B gene in progressive myoclonus epilepsy.** *Nature* **386**, 847-851.

Lalioti, M. D., Scott, H. S., Genton, P., Grid, D., Ouazzani, R., M'Rabet, A., Ibrahim, S., Gouider, R., Dravet, C., Chkili, T., Bottani, A., Buresi, C., Malafosse, A. & Antonarakis, S. E. (1998). **A PCR amplification method reveals instability of the dodecamer repeat in**

progressive myoclonus epilepsy (EPM1) and no correlation between the size of the repeat and age at onset. *Am. J. Hum. Genet.* **62**, 842-847.

Lavedan, C., Grabczyk, E., Usdin, K. & Nussbaum, R. L. (1998). **Long uninterrupted CGG repeats within the first exon of the human FMR1 gene are not intrinsically unstable in transgenic mice.** *Genomics* **50**, 229-240.

Leach, D. R. F. (1994). **Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair.** *BioEssays* **16**, 893-900.

Ledbetter, D. H., Ledbetter, S. A. & Nussbaum, R. L. (1986). **Implications of fragile-X expression in normal males for the nature of the mutation.** *Nature* **324**, 161-163.

Lee, J. S. (1990). **The stability of polypurine tetraplexes in the presence of mono- and divalent cations.** *Nucl. Acids Res.* **18**, 6,057-6,060.

Lee, J. S., Evans, D. H. & Morgan, A. R. (1980). **Polypurine DNAs and RNAs form secondary structures which may be tetra-stranded.** *Nucl. Acids Res.* **8**, 4,305-4,320.

Leggo, J., Dalton, A., Morrison, P. J., Dodge, A., Connarty, M., Kotze, M. J. & Rubinsztein, D. C. (1997). **Analysis of spinocerebellar ataxia types 1, 2, 3, and 6, dentatorubral-pallidoluysian atrophy, and Friedreich's ataxia genes in spinocerebellar ataxia patients in the UK.** *J. Med. Genet.* **34**, 982-985.

Leonard, G. A., Zhang, S., Peterson, M. R., Harrop, S. J., Helliwell, J. R., Cruse, W. B. T., Langlois d'Estaintot, B., Kennard, O., Brown, T. & Hunter, W. N. (1995). **Self-association of a DNA loop creates a quadruplex: crystal structure of d(GCATGCT) at 1.8 Å resolution.** *Structure* **3**, 335-340.

Leroy, J.-L., Gehring, K., Kettani, A. & Guéron, M. (1993). **Acid multimers of oligodeoxycytidine strands: stoichiometry, base-pair characterization, and proton exchange properties.** *Biochemistry* **32**, 6,019-6,031.

Levinson, G. & Gutman, G. A. (1987). **Slipped-strand mispairing: a major mechanism for DNA sequence evolution.** *Mol. Biol. Evol.* **4**, 203-221.

Lilley, D. M. J. (1985). **The kinetic properties of cruciform extrusion are determined by DNA base-sequence.** *Nucl. Acids Res.* **13**, 1,443-1,465.

Limprasert, P., Nouri, N., Heyman, R. A., Nopparatana, C., Kamonsilp, M., Deininger, P. L. & Keats, B. J. B. (1996). **Analysis of CAG repeat of the Machado-Joseph gene in human, chimpanzee and monkey populations: a variant nucleotide is associated with the number of CAG repeats.** *Hum. Mol. Genet.* **5**, 207-213.

Lindblad, K., Zander, C., Schalling, M. & Hudson, T. (1994). **Growing triplet repeats.** *Nature Genet.* **7**, 124.

Lindblom, A., Tannergård, P., Werelius, B. & Nordenskjöld, M. (1993). **Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer.** *Nature Genet.* **5**, 279-282.

Liskay, R. M., Letson, A. & Stachelek, J. (1987). **Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells.** *Genetics* **115**, 161-167.

Lloyd, R. G. & Buckman, C. (1985). **Identification and genetic analysis of *sbcC* mutations in commonly used *recBC sbcB* strains of *Escherichia coli* K-12.** *J. Bacteriol.* **164**, 836-844.

Logan, C., Willard, H. F., Rommens, J. M. & Joyner, A. L. (1989). **Chromosomal localization of the human homeo box-containing genes, EN1 and EN2.** *Genomics* **4**, 206-209.

Longo, F. M. & Massa, S. M. (1997). **Trinucleotide repeats in transgenic mice: new insights.** *The Neuroscientist* **3**, 273-275.

Longshore, J. W. & Tarleton, J. (1996). **Dynamic mutations in human genes: a review of trinucleotide repeat diseases.** *J. Genet.* **75**, 193-217.

Lorenzetti, D., Bohlega, S. & Zoghbi, H. Y. (1997). **The expansion of the CAG repeat in ataxin-2 is a frequent cause of autosomal dominant spinocerebellar ataxia.** *Neurology* **49**, 1,009-1,013.

Löwdin, P.-O. (1964). **Some aspects on DNA replication; incorporation errors and proton transfer.** In *Electronic aspects of biochemistry* (Pullman, B., ed.), Academic Press, New York and London, pp. 167-201.

Lubs, H. A. (1969). **A marker X chromosome.** *Am. J. Hum. Genet.* **21**, 231-244.

Macdonald, M. E., Ambrose, C. M., Duyao, M. P., Myers, R. H., Lin, C., Srinidhi, L., Barnes, G., Taylor, S. A., James, M., Groot, N., Macfarlane, H., Jenkins, B., Anderson, M. A., Wexler, N. S., Gusella, J. F., Bates, G. P., Baxendale, S., Hummerich, H., Kirby, S., North, M., Youngman, S., Mott, R., Zehetner, G., Sedlacek, Z., Poustka, A., Frischauf, A.-M., Lehrach, H., Buckler, A. J., Church, D., Doucette-Stamm, L., O'Donovan, M. C., Riba-Ramirez, L., Shah, M., Stanton, V. P., Strobel, S. A., Draths, K. M., Wales, J. L.,

Dervan, P., Housman, D. E., Altherr, M., Shiang, R., Thompson, L., Fielder, T., Wasmuth, J. J., Tagle, D., Valdes, J., Elmer, L., Allard, M., Castilla, L., Swaroop, M., Blanchard, K., Collins, F. S., Snell, R., Holloway, T., Gillespie, K., Datson, N., Shaw, D. & Harper, P. S. (1993). **A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes.** *Cell* **72**, 971-983.

Macpherson, J., Harvey, J., Curtis, G., Webb, T., Heitz, D., Rousseau, F. & Jacobs, P. (1992). **A reinvestigation of thirty three fragile(X) families using probe StB12.3.** *Am. J. Med. Genet.* **43**, 905-912.

Maddox, J. (1994). **Triplet repeat genes raise questions.** *Nature* **368**, 685.

Mahadevan, M. S., Foitzik, M. A., Surh, L. C. & Korneluk, R. G. (1993). **Characterization and polymerase chain reaction (PCR) detection of an *Alu* deletion polymorphism in total linkage disequilibrium with myotonic dystrophy.** *Genomics* **15**, 446-448.

Mahadevan, M., Tsilfidis, C., Sabourin, L., Shutler, G., Amemiya, C., Jansen, G., Neville, C., Narang, M., Barcelo, J., O'Hoy, K., Leblond, S., Earle-Macdonald, J., de Jong, P. J., Weiringa, B. & Korneluk, R. G. (1992). **Myotonic dystrophy mutation: an unstable CTG repeat in the 3' untranslated region of the gene.** *Science* **255**, 1,253-1,255.

Mahtani, M. M. & Willard, H. F. (1993). **A polymorphic X-linked tetranucleotide repeat locus displaying a high rate of new mutation: implications for mechanisms of mutation at short tandem repeat loci.** *Hum. Mol. Genet.* **2**, 431-437.

Malter, H. E., Iber, J. C., Willemsen, R., de Graaff, E., Tarleton, J. C., Leisti, J., Warren, S. T. & Oostra, B. A. (1997). **Characterization of the full fragile X syndrome mutation in fetal gametes.** *Nature Genet.* **15**, 165-169.

Mandel, J.-L. (1997). **Breaking the rule of three.** *Nature* **386**, 767-769.

Maniatis, T., Jeffrey, A. & van deSande, H. (1975). **Chain length determination of small double- and single-stranded DNA molecules by polyacrylamide gel electrophoresis.** *Biochemistry* **14**, 3,787-3,794.

Mariappan, S. V. S., Catasti, P., Chen, X., Ratliff, R., Moysis, R. K., Bradbury, E. M. & Gupta, G. (1996b). **Solution structures of the individual single strands of the fragile X DNA triplets $(GCC)_n$·$(GGC)_n$.** *Nucl. Acids Res.* **24**, 784-792.

Mariappan, S. V. S., Catasti, P., Silks, L. A., III, Bradbury, E. M. & Gupta, G. (1999). **The high-resolution structure of the triplex formed by the GAA/TTC triplet repeat associated with Friedreich's ataxia.** *J. Mol. Biol.* **285**, 2,035-2,052.

Mariappan, S. V. S., Garcia, A. E. & Gupta, G. (1996a). **Structure and dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy.** *Nucl. Acids Res.* **24**, 775-783.

Mariappan, S. V. S., Silks, L. A., III, Chen, X., Springer, P. A., Wu, R., Moyzis, R. K., Bradbury, E. M., Garcia, A. E. & Gupta, G. (1998a). **Solution structures of the Huntington's disease DNA triplets, $(CAG)_n$.** *J. Biomol. Struct. Dyn.* **15**, 723-744.

Mariappan, S. V. S., Silks, L. A., III, Bradbury, E. M. & Gupta, G. (1998b). **Fragile X DNA triplet repeats, $(GCC)_n$, form hairpins with single hydrogen-bonded cytosine·cytosine mispairs at the CpG sites: isotope-edited nuclear magnetic resonance spectroscopy on $(GCC)_n$ with selective ^{15}N4-labeled cytosine bases.** *J. Mol. Biol.* **283**, 111-120.

Martin, J. P. & Bell, J. (1943). **A pedigree of mental defect showing sex-linkage.** *J. Neurol. Psych.* **6**, 154-157.

Matilla, T., Volpini, V., Genís, D., Rosell, J., Corral, J., Dávalos, A., Molins, A. & Estivill, X. (1993). **Presymptomatic analysis of spinocerebellar ataxia type 1 (SCA1) via the expansion of the SCA1 CAG-repeat in a large pedigree displaying anticipation and parental male bias.** *Hum. Mol. Genet.* **2**, 2,123-2,128.

Matsumura, R., Takayanagi, T., Murata, K., Futamura, N., Hirano, M. & Ueno, S. (1996). **Relationship of (CAG)nC configuration to repeat instability of the Machado-Joseph disease gene.** *Hum. Genet.* **98**, 643-645.

Matsuyama, Z., Kawakami, H., Maruyama, H., Izumi, Y., Komure, O., Udaka, F., Kameyama, M., Nishio, T., Kuroda, Y., Nishimura, M. & Nakamura, S. (1997). **Molecular features of the CAG repeats of spinocerebellar ataxia 6 (SCA6).** *Hum. Mol. Genet.* **6**, 1,283-1,287.

Mäueler, W., Kyas, A., Keyl, H.-G. & Epplen, J. T. (1998). **A genome-derived $(gaa.ttc)_{24}$ trinucleotide block binds nuclear protein(s) specifically and forms triple helices.** *Gene* **215**, 389-403.

Maurer, D. J., O'Callaghan, B. L. & Livingston, D. M. (1996). **Orientation dependence of trinucleotide CAG repeat instability in *Saccharomyces cerevisiae*.** *Mol. Cell. Biol.* **16**, 6,617-6,622.

Maxam, A. M. & Gilbert, W. (1977). **A new method for sequencing DNA.** *Proc. Natl Acad. Sci. U.S.A.* **74**, 560-564.

McGavin, S. (1971). **Models of specifically paired like (homologous) nucleic acid structures.** *J. Mol. Biol.* **55**, 293-298.

McInnis, M. G. (1996). **Anticipation: an old idea in new genes.** *Am. J. Hum. Genet.* **59**, 973-979.

McMurray, C. T. (1995). **Mechanisms of DNA expansion.** *Chromosoma* **104**, 2-13.

McNeil, S. M., Novelletto, A., Srinidhi, J., Barnes, G., Kornbluth, I., Altherr, M. R., Wasmuth, J. J., Gusella, J. F., MacDonald, M. E. & Myers, R. H. (1997). **Reduced penetrance of the Huntington's disease mutation.** *Hum. Mol. Genet.* **6**, 775-779.

Michaelis, R. C., Velagaleti, G. V., Jones, C., Pivnick, E. K., Phelan, M. C., Boyd, E., Tarleton, J., Wilroy, R. S., Tunnacliffe, A. & Tharapel, A. T. (1998). **Most Jacobsen syndrome deletion breakpoints occur distal to *FRA11B*.** *Am. J. Med. Genet.* **76**, 222-228.

Miret, J. J., Pessoa-Brandão, L. & Lahue, R. S. (1997). **Instability of CAG and CTG trinucleotide repeats in *Saccharomyces cerevisiae*.** *Mol. Cell. Biol.* **17**, 3,382-3,387.

Miret, J. J., Pessoa-Brandão, L. & Lahue, R. S. (1998). **Orientation-dependent and sequence-specific expansions of CTG/CAG trinucleotide repeats in *Saccharomyces cerevisiae*.** *Proc. Natl Acad. Sci. U.S.A.* **95**, 12,438-12,443.

Mishmar, D., Rahat, A., Scherer, S. W., Nyakatura, G., Hinzmann, B., Kohwi, Y., Mandel-Gutfroind, Y., Lee, J. R., Drescher, B., Sas, D. E., Margalit, H., Platzer, M., Weiss, A., Tsui, L.-C., Rosenthal, A. & Kerem, B. (1998). **Molecular characterization of a common fragile site (*FRA7H*) on human chromosome 7 by the cloning of a simian virus 40 integration site.** *Proc. Natl Acad. Sci. U.S.A.* **95**, 8141-6.

Mitas, M. (1997). **Trinucleotide repeats associated with human disease.** *Nucl. Acids Res.* **25**, 2,245-2,254.

Mitas, M., Yu, A., Dill, J. & Haworth, I. S. (1995b). **The trinucleotide repeat sequence d(CGG)$_{15}$ forms a heat-stable hairpin containing G^{syn}·G^{anti} base pairs.** *Biochemistry* **34**, 12,803-12,811.

Mitas, M., Yu, A., Dill, J., Kamp, T. J., Chambers, E. J. & Haworth, I. S. (1995a). **Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: (CTG)$_{15}$.** *Nucleic Acids Res.* **23**, 1,050-1,059.

Mitchell, J. E., Newbury, S. F. & McClellan, J. A. (1995). **Compact structures of d(CNG)$_n$ oligonucleotides in solution and their possible relevance to Fragile X and related human genetic diseases.** *Nucl. Acids Res.* **23**, 1,876-1,881.

Monckton, D. G., Neumann, R., Guram, T., Fretwell, N., Tamaki, K., MacLeod, A. & Jeffreys, A. J. (1994). **Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism.** *Nature Genet.* **8**, 162-170.

Morton, N. E. & Macpherson, J. N. (1992). **Population genetics of the fragile-X syndrome: multiallelic model for the *FMR1* locus.** *Proc. Natl Acad. Sci. U.S.A.* **89**, 4,215-4,217.

Moutou, C., Vincent, M. C., Biancalana, V. & Mandel, J.-L. (1997). **Transition from premutation to full mutation in fragile X syndrome is likely to be prezygotic.** *Hum. Mol. Genet.* **6**, 971-979.

Mundlos, S., Otto, F., Mundlos, C., Mulliken, J. B., Aylsworth, A. S., Albright, S., Lindhout, D., Cole, W. G., Henn, W., Knoll, J. H., Owen, M. J., Mertelsmann, R., Zabel, B. U. & Olsen, B. R. (1997). **Mutations involving the transcription factor CBFA1 cause cleidocranial dysplasia.** *Cell* **89**, 773-779.

Muragaki, Y., Mundlos, S., Upton, J. & Olsen, B. R. (1996). **Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13.** *Science* **272**, 548-551.

Murchie, A. I. H., Bowater, R., Aboul-ela, F. & Lilley, D. M. J. (1992). **Helix opening transitions in supercoiled DNA.** *Biochim. Biophys. Acta* **1131**, 1-15.

Murray, J., Buard, J., Neil, D. L., Yeramian, E., Tamaki, K., Hollies, C. & Jeffreys, A. J. (1999). **Comparative Sequence Analysis of Human Minisatellites Showing Meiotic Repeat Instability.** *Genome Res* **9**, 130-136.

Mytelka, D. S. & Chamberlin, M. J. (1996). **Analysis and suppression of DNA polymerase pauses associated with a trinucleotide consensus.** *Nucl. Acids Res.* **24**, 2,774-2,781.

Nadel, Y., Weisman-Shomer, P. & Fry, M. (1995). **The fragile X syndrome single strand d(CGG)$_n$ nucleotide repeats readily fold back to form unimolecular hairpin structures.** *J. Biol. Chem.* **270**, 28,970-28,977.

Nagafuchi, S., Yanagisawa, H., Sato, K., Shirayama, T., Ohsaki, E., Bundo, M., Takeda, T., Tadokoro, K., Kondo, I., Murayama, N., Tanaka, Y., Kikushima, H., Umino, K., Kurosawa, H., Furukawa, T., Nihei, K., Inoue, T., Sano, A., Komure, O., Takahashi, M., Yoshizawa, T., Kanazawa, I. & Yamada, M. (1994). **Dentatorubral and pallidoluysian**

atrophy expansion of an unstable CAG trinucleotide on chromosome 12p. *Nature Genet.* **6**, 14-18.

Nakamoto, M., Takebayashi, H., Kawaguchi, Y., Narumiya, S., Taniwaki, M., Nakamura, Y., Ishikawa, Y., Akiguchi, I., Kimura, J. & Kakizuka, A. (1997). **A CAG/CTG expansion in the normal population.** *Nature Genet.* **17**, 385-386.

Nancarrow, J. K., Holman, K., Mangelsdorf, M., Hori, T., Denton, M., Sutherland, G. R. & Richards, R. I. (1995). **Molecular basis of p(CCG)n repeat instability at the *FRA16A* fragile site locus.** *Hum. Mol. Genet.* **4**, 367-372.

Nancarrow, J. K., Kremer, E., Holman, K., Eyre, H., Dogget, N. A., LePaslier, D., Callen, D. F., Sutherland, G. R. & Richards, R. I. (1994). **Implications of *FRA16A* structure for the mechanism of chromosomal fragile site genesis.** *Science* **264**, 1,938-1,941.

Nasmyth, K. A. (1982). **Molecular genetics of yeast mating type.** *Ann. Rev. Genet.* **16**, 439-500.

Neville, C. E., Mahadevan, M. S., Barceló, J. M. & Korneluk, R. G. (1994). **High resolution genetic analysis suggests one ancestral predisposing haplotype for the origin of the myotonic dystrophy mutation.** *Hum. Mol. Genet.* **3**, 45-51.

Nicolaides, N. C., Papadopoulos, N., Liu, B., Wei, Y.-F., Carter, K. C., Ruben, S. M., Rosen, C. A., Haseltine, W. A., Fleischmann, R. D., Fraser, C. M., Adams, M. D., Venter, J. C., Dunlop, M. G., Hamilton, S. R., Petersen, G. M., de la Chapelle, A., Vogelstein, B. & Kinzler, K. W. (1994). **Mutations of two *PMS* homologues in hereditary nonpolyposis colon cancer.** *Nature* **371**, 75-80.

Nürnberg, P., Roewer, L., Neitzel, H., Sperling, K., Pöpperl, A., Hundrieser, J., Pöche, H., Epplen, C., Zishler, H. & Epplen, J. T. (1989). **DNA fingerprinting with the oligonucleotide probe $(CAC)_5/(GTG)_5$: somatic stability and germline mutations.** *Hum. Genet.* **84**, 75-78.

Nussbaum, R. L., Airhart, S. D. & Ledbetter, D. H. (1986). **Recombination and amplification of pyrimidine-rich sequences may be responsible for initiation and progression of the Xq27 fragile site: a hypothesis.** *Am. J. Med. Genet.* **23**, 715-721.

O'Brien, E. J. (1967). **Crystal structures of two complexes containing guanine and cytosine derivatives.** *Acta Cryst.* **23**, 92-106.

O'Donovan, M. C., Guy, C., Craddock, N., Bowen, T., McKeon, P., Macedo, A., Maier, W., Wildenauer, D., Aschauer, H. N., Sorbi, S., Feldman, E., Mynett-Johnson, L., Claffey, E., Nacmias, B., Valente, J., Dourado, A., Grassi, E., Lenzinger, E., Heiden, A. M., Moorhead, S., Harrison, D., Williams, J., McGuffin, P. & Owen, M. J. (1996a). **Confirmation of association between expanded CAG/CTG repeats and both schizophrenia and bipolar disorder.** *Psychol. Med.* **26**, 1,145-1,153.

O'Donovan, M. C., Craddock, N., Guy, C., McGuffin, P. & Owen, M. (1996b). **Involvement of expanded trinucleotide repeats in common diseases.** *Lancet* **348**, 1,739-1,740.

O'Hoy, K. L., Tsilfidis, C., Mahadevan, M. S., Neville, C. E., Barcelló, J., Hunter, A. G. W. & Korneluk, R. G. (1993). **Reduction in size of the myotonic dystrophy trinucleotide repeat mutation during transmission.** *Science* **259**, 809-812.

Oberlé, I., Heilig, R., Moisan, J. P., Kloepfer, C., Mattéi, G. M., Mattéi, J. F., Boué, J., Froster-Iskenius, U., Jacobs, P. A., Lathrop, G. M., Lalouel, J. M. & Mandel, J.-L. (1986). **Genetic analysis of the fragile-X mental retardation syndrome with two flanking polymorphic DNA markers.** *Proc. Natl Acad. Sci. U.S.A.* **83**, 1,016-1,020.

Oberlé, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boue, J., Bertheas, M. F. & Mandel, J. L. (1991). **Instability of a 550-Base Pair DNA Segment and Abnormal Methylation In Fragile X Syndrome.** *Science* **252**, 1,097-1,102.

Oda, T., Kitamoto, T., Tateishi, J., Mitsuhashi, T., Iwabuchi, K., Haga, C., Oguni, E., Kato, Y., Tominaga, I., Yanai, K., Kashima, H., Kogure, T., Hori, K. & Ogino, K. (1995). **Prion disease with 144-base-pair insertion in a Japanese family line.** *Acta Neuropathol (Berl)* **90**, 80-86.

Ohshima, K., Kang, S., Larson, J. E. & Wells, R. D. (1996a). **Cloning, characterization, and properties of seven triplet repeat DNA sequences.** *J. Biol. Chem.* **271**, 16,773-16,783.

Ohshima, K., Kang, S., Larson, J. E. & Wells, R. D. (1996b). **TTA·TAA triplet repeats in plasmids form a non-H bonded structure.** *J. Biol. Chem.* **271**, 16,784-16,791.

Ohshima, K., Kang, S. & Wells, R. D. (1996). **CTG triplet repeats from human hereditary diseases are dominant genetic expansion products in *Escherichia coli*.** *J. Biol. Chem.* **271**, 1,853-1,856.

Ohshima, K. & Wells, R. D. (1997). **Hairpin formation during DNA synthesis primer realignment *in vitro* in triplet repeat sequences from human hereditary disease genes.** *J. Biol. Chem.* **272**, 16,798-16,806.

Parker, B. O. & Marinus, M. G. (1992). **Repair of DNA heteroduplexes containing small heterologous sequences in *Escherichia coli*.** *Proc. Natl Acad. Sci. U.S.A.* **89**, 1,730-1,734.

Parrish, J. E., Oostra, B. A., Verkerk, A. J. M. H., Richards, C. S., Reynolds, J., Spikes, A. S., Shaffer, L. G. & Nelson, D. L. (1994). **Isolation of a GCC repeat showing expansion in FRAXF, a fragile site distal to FRAXA and FRAXE.** *Nature Genet.* **8**, 229-235.

Paull, T. T. & Gellert, M. (1998). **The 3' to 5' exonuclease activity of Mre 11 facilitates repair of DNA double-strand breaks.** *Mol. Cell* **1**, 969-979.

Pearson, C. E., Eichler, E. E., Lorenzetti, D., Kramer, S. F., Zoghbi, H. Y., Nelson, D. L. & Sinden, R. R. (1998b). **Interruptions in the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation.** *Biochemistry* **37**, 2,701-2,708.

Pearson, C. E., Ewel, A., Acharya, S., Fishel, R. A. & Sinden, R. R. (1997). **Human MSH2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases.** *Hum. Mol. Genet.* **6**, 1,117-1,123.

Pearson, C. E. & Sinden, R. R. (1996). **Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci.** *Biochemistry* **35**, 5,041-5,053.

Pearson, C. E., Wang, Y.-H., Griffith, J. D. & Sinden, R. R. (1998a). **Structural analysis of slipped-strand DNA (S-DNA) formed in $(CTG)_n \cdot (CAG)_n$ repeats from the myotonic dystrophy locus.** *Nucl. Acids Res.* **26**, 816-823.

Peltomäki, P., Aaltonen, L. A., Sistonen, P., Pylkkänen, L., Mecklin, J.-P., Järvinen, H., Green, J. S., Jass, J. R., Weber, J. L., Leach, F. S., Petersen, G. M., Hamilton, S. R., de la Chapelle, A. & Vogelstein, B. (1993). **Genetic mapping of a locus predisposing to human colorectal cancer.** *Science* **260**, 810-812.

Pembrey, M. E. & Winter, R. M. (1985). **Further segregation analysis of the fragile X syndrome with special reference to transmitting males.** *Human Genetics* **71**, 182.

Pembrey, M. E., Winter, R. M. & Davies, K. E. (1984). **A premutation that generates the definitive mutation by recombination explains the inheritance of the Martin-Bell syndrome (fragile X).** *J. Med. Genet.* **21**, 299-299.

Pembrey, M. E., Winter, R. M. & Davies, K. E. (1985). **A premutation that generates a defect at crossing over explains the inheritance of fragile X mental retardation.** *Am. J. Med. Genet.* **21**, 709-717.

Penrose, L. S. (1948). **The problem of anticipation in pedigrees of dystrophia myotonica.** *Ann. Eugen.* **14**, 125-132.

Petes, T. D., Greenwell, P. W. & Dominska, M. (1997). **Stabilization of microsatellite sequences by variant repeats in the yeast *Saccharomyces cerevisiae*.** *Genetics* **146**, 491-498.

Petruska, J., Arnheim, N. & Goodman, M. F. (1996). **Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases.** *Nucl. Acids Res.* **24**, 1,992-1,998.

Petruska, J., Hartenstine, M. J. & Goodman, M. F. (1998). **Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease.** *J. Biol. Chem.* **273**, 5,204-5,210.

Pieretti, M., Zhang, F. P., Fu, Y. H., Warren, S. T., Oostra, B. A., Caskey, C. T. & Nelson, D. L. (1991). **Absence of expression of the *FMR-1* gene in fragile X syndrome.** *Cell* **66**, 817-822.

Pinder, D. J., Blake, C. E., Lindsey, J. C. & Leach, D. R. (1998). **Replication strand preference for deletions associated with DNA palindromes.** *Mol. Microbiol.* **28**, 719-727.

Pribnow, D., Sigurdson, D. C., Gold, L., Singer, B. S., Napoli, C., Brosius, J., Dull, T. J. & Noller, H. F. (1981). **rII cistrons of bacteriophage T4: DNA sequence around the intercistronic divide and positions of genetic landmarks.** *J. Mol. Biol.* **149**, 337-376.

Pulst, S.-M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X.-N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A., DeJong, P., Rouleau, G. A., Auburger, G., Korenberg, J. R., Figueroa, C. & Sahba, S. (1996). **Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2.** *Nature Genet.* **14**, 269-276.

Reddy, P. S. & Housman, D. E. (1997). **The complex pathology of trinucleotide repeats.** *Curr. Opin. Cell Biol.* **9**, 364-372.

Réfrégiers, M., Laigle, A., Jollès, B., Wheeler, G. V. & Chinsky, L. (1997). **Resonance Raman analysis of a fluorescently labeled oligonucleotide forming a very stable hairpin.** *Eur. Biophys. J. with Biophys. Letters* **26**, 277-281.

Resnick, M. (1976). **The repair of double strand breaks in DNA: a model involving recombination.** *J. theor. Biol.* **59**, 97-106.

Richard, G.-F. & Dujon, B. (1996). **Distribution and variability of trinucleotide repeats in the genome of the yeast *Saccharomyces cerevisiae*.** *Gene* **174**, 165-174.

Richards, R. I., Holman, K., Yu, S. & Sutherland, G. R. (1993). **Fragile-X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding-sites for specific nuclear proteins.** *Hum. Mol. Genet.* **2**, 1,429-1,435.

Richards, R. I. & Sutherland, G. R. (1992a). **Heritable unstable DNA sequences.** *Nature Genet.* **1**, 7-9.

Richards, R. I. & Sutherland, G. R. (1992b). **Dynamic mutations: a new class of mutations causing human disease.** *Cell* **70**, 709-712.

Richards, R. I. & Sutherland, G. R. (1994). **Simple repeat DNA is not replicated simply.** *Nature Genetics* **6**, 114-116.

Richards, R. I. & Sutherland, G. R. (1997). **Dynamic mutation: possible mechanisms and significance in human disease.** *Trends Biochem. Sci.* **22**, 432-436.

Riess, O., Schöls, L., Böttger, H., Nolte, D., Vieira-Saecker, A. M., Schimming, C., Kreuz, F., Macek, M., Jr., Krebsova, A., Macek, M. S., Klockgether, T., Zühlke, C. & Laccone, F. A. (1997). **SCA6 is caused by moderate CAG expansion in the α_{1A}-voltage-dependent calcium channel gene.** *Hum. Mol. Genet.* **6**, 1,289-1,293.

Ripley, L. S. (1982). **Model for the participation of quasi-palindromic DNA sequences in frameshift mutation.** *Proc. Natl Acad. Sci. U.S.A.* **79**, 4,128-4,132.

Ritchie, R. J., Chakrabarti, L., Knight, S. J. L., Harding, R. M. & Davies, K. E. (1997). **Population genetics of the *FRAXE* and *FRAXF* GCC repeats, and a novel CGG repeat, in Xq28.** *Am. J. Med. Genet.* **73**, 463-469.

Ritchie, R. J., Knight, S. J. L., Hirst, M. C., Grewal, P. K., Bobrow, M., Cross, G. S. & Davies, K. E. (1994). **The cloning of *FRAXF*: trinucleotide repeat expansion and methylation at a third fragile site in distal Xqter.** *Hum. Mol. Genet.* **3**, 2,115-2,121.

Rosche, W. A., Jaworski, A., Kang, S., Kramer, S. F., Larson, J. E., Geidroc, D. P., Wells, R. D. & Sinden, R. R. (1996). **Single-stranded DNA-binding protein enhances the stability of CTG triplet repeats in *Escherichia coli*.** *J. Bacteriol.* **178**, 5,042-5,044.

Rosche, W. A., Trinh, T. Q. & Sinden, R. R. (1995). **Differential DNA secondary structure-mediated deletion mutation in the leading and lagging strands.** *J. Bacteriol.* **177**, 4,385-4,391.

Rotwein, P., Yokoyama, S., Didier, D. K. & Chirgwin, J. M. (1986). **Genetic analysis of the hypervariable region flanking the human insulin gene.** *Am. J. Hum. Genet.* **39**, 291-299.

Rubinsztein, D. C., Barton, D. E., Davison, B. C. C. & Ferguson-Smith, M. A. (1993a). **Analysis of the huntingtin gene reveals a trinucleotide-length polymorphism in the region of the gene that contains two CCG-rich stretches and a correlation between decreased age of onset of Huntington's disease and CAG repeat number.** *Hum. Mol. Genet.* **2**, 1,713-1,715.

Rubinsztein, D. C., Leggo, J., Amos, W., Barton, D. E. & Ferguson-Smith, M. A. (1994). **Myotonic dystrophy CTG repeats and the associated insertion/deletion polymorphism in human and primate populations.** *Hum. Mol. Genet.* **3**, 2,031-2,035.

Rubinsztein, D. C., Leggo, J., Barton, D. E. & Ferguson-Smith, M. A. (1993b). **Site of (CCG) polymorphism in the HD gene.** *Nature Genet.* **6**, 214-215.

Rubinsztein, D. C., Leggo, J., Coles, R., Almqvist, E., Biancalana, V., Cassiman, J.-J., Chotai, K., Connarty, M., Craufurd, D., Curtis, A., Curtis, D., Davidson, M. J., Differ, A.-M., Dode, C., Dodge, A., Frontali, M., Ranen, N. G., Stine, O. C., Sherr, M., Abbott, M. H., Franz, M. L., Graham, C. A., Harper, P. S., Hedreen, J. C., Jackson, A., Kaplan, J.-C., Losekoot, M., Macmillan, J. C., Morrison, P., Trottier, Y., Novelletto, A., Simpson, S. A., Theilmann, J., Whittaker, J. L., Folstein, S. E., Ross, C. A. & Hayden, M. R. (1996). **Phenotypic characterization of individuals with 30-40 CAG repeats in the Huntington Disease (HD) gene reveals HD cases with 36 repeats and apparently normal elderly individuals with 36-39 repeats.** *Am. J. Hum. Genet.* **59**, 16-22.

Rubnitz, J. & Subramani, S. (1984). **The minimum amount of homology required for homologous recombination in mammalian cells.** *Mol. Cell. Biol.* **4**, 2,253-2,258.

Samadashwily, G. M., Raca, G. & Mirkin, S. M. (1997). **Trinucleotide repeats affect DNA replication *in vivo*.** *Nature Genet.* **17**, 298-304.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*. 2nd edit (Irwin, N., Ed.), Cold Spring Harbor Press, U.S.A.

Sanpei, K., Takano, H., Igarashi, S., Sato, T., Oyake, M., Sasaki, H., Wakisaka, A., Tashiro, K., Ishida, Y., Ikeuchi, T., Koide, R., Saito, M., Sato, A., Tanaka, T., Hanyu, S., Takiyama, Y., Nishizawa, M., Shimizu, N., Nomura, Y., Segawa, M., Iwabuchi, K., Eguchi, I., Tanaka, H., Takahashi, H. & Tsuji, S. (1996). **Identification of the spinocerebellar ataxia type 2 gene using a direct identification of repeat expansion and cloning technique, DIRECT.** *Nature Genet.* **14**, 277-284.

Sarkar, P. S., Chang, H.-C., Boudi, F. B. & Reddy, S. (1998). **CTG repeats show bimodal amplification in *E. coli*.** *Cell* **95**, 531-540.

Sato, K., Kashihara, K., Okada, S., Ikeuchi, T., Tsuji, S., Shomori, T., Morimoto, K. & Hayabara, T. (1995). **Does homozygosity advance the onset of dentatorubral-pallidoluysian atrophy?** *Neurology* **45**, 1,934-1,936.

Schäfer, R., Zischler, H., Birsner, U., Becker, A. & Epplen, J. T. (1988). **Optimized oligonucleotide probes for DNA fingerprinting.** *Electrophoresis* **9**, 369-374.

Schalling, M., Hudson, T. J., Buetow, K. H. & Housman, D. E. (1993). **Direct detection of novel expanded trinucleotide repeats in the human genome.** *Nature Genet.* **4**, 135-139.

Schellman, J. A. (1974). **Flexibility of DNA.** *Biopolymers* **13**, 217-226.

Schlötterer, C. & Tautz, D. (1992). **Slippage synthesis of simple sequence DNA.** *Nucl. Acids Res.* **20**, 211-215.

Schumacher, S., Fuchs, R. P. P. & Bichara, M. (1998). **Expansion of CTG repeats from human disease genes is dependent upon replication mechanisms in *Escherichia coli*: the effect of long patch mismatch repair revisited.** *J. Mol. Biol.* **279**, 1,101-1,110.

Schweitzer, J. K. & Livingston, D. M. (1997). **Destabilization of CAG trinucleotide repeat tracts by mismatch repair mutations in yeast.** *Hum. Mol. Genet.* **6**, 349-355.

Sharples, G. J. & Leach, D. R. F. (1995). **Structural and functional similarities between the SbcCD proteins of *Escherichia coli* and the RAD50 and MRE11 (RAD32) recombination and repair proteins of yeast.** *Mol. Microbiol.* **17**, 1,215-1,217.

Shen, P. & Huang, H. V. (1986). **Homologous recombination in *Escherichia coli*: dependence on length and homology.** *Genetics* **112**, 441-457.

Sherman, S. L., Jacobs, P. A. & Morton, N. E. (1985b). **Further segregation analysis of the fragile X-Syndrome with special reference to transmitting males - Reply.** *Hum. Genet.* **71**, 183.

Sherman, S. L., Jacobs, P. A., Morton, N. E., Frosteriskenius, U., Howardpeebles, P. N., Nielsen, K. B., Partington, M. W., Sutherland, G. R., Turner, G. & Watson, M. (1985a). **Further segregation analysis of the fragile-X syndrome with special reference to transmitting males.** *Hum. Genet.* **69**, 289-299.

Sherman, S. L., Morton, N. E., Jacobs, P. A. & Turner, G. (1984). **The marker (X) syndrome: a cytogenetic and genetic analysis.** *Ann. Hum. Genet.* **48**, 21-37.

Shibata, D., Peinado, M. A., Ionov, Y., Malkhosyan, S. & Perucho, M. (1994). **Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation.** *Nature Genet.* **6**, 273-281.

Shimizu, M., Gellibolian, R., Oostra, B. A. & Wells, R. D. (1996). **Cloning, characterization and properties of plasmids containing CGG triplet repeats from the FMR-1 gene.** *J. Mol. Biol.* **258**, 614-626.

Shimizu, M., Hanvey, J. C. & Wells, R. D. (1989). **Intramolecular DNA triplexes in supercoiled plasmids. I. Effect of loop size on formation and stability.** *J. Biol. Chem.* **264**, 5,944-5,949.

Sen, D. & Gilbert, W. (1990). **A sodium-potassium switch in the formation of four-stranded G4-DNA.** *Nature* **344**, 410-414.

Siedlaczck, I., Epplen, C., Riess, O. & Epplen, J. T. (1993). **Simple repetitive (GAA)n loci in the human genome.** *Electrophoresis* **14**, 973-977.

Silva, A. J., Johnson, J. P. & White, R. L. (1987). **Characterization of a highly polymorphic region 5' to JH in the human immunoglobulin heavy chain.** *Nucl. Acids Res.* **15**, 2845-2857.

Sinden, R. R. & Wells, R. D. (1992). **DNA structure, mutations and human genetic disease.** *Cur. Opin. Biotechnol.* **3**, 612-622.

Slightom, J. L., Blechl, A. E. & Smithies, O. (1980). **Human fetal G gamma- and A gamma-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes.** *Cell* **21**, 627-638.

Smith, G. K., Jie, J., Fox, G. E. & Gao, X. (1995). **DNA CTG triplet repeats involved in dynamic mutations of neurologically related gene sequences form stable duplexes.** *Nucl. Acids Res.* **23**, 4,303-4,311.

Smith, S. S. (1991). **DNA methylation in eukaryotic chromosome stability.** *Mol. Carcinog.* **4**, 91-92.

Smith, S. S. & Baker, D. J. (1997). **Stalling of human methyltransferase at single-strand conformers from the Huntington's locus.** *Biochem. Biophys. Res. Commun.* **234**, 73-78.

Smith, S. S., Hardy, T. A. & Baker, D. J. (1987). **Human DNA (cytosine-5)methyltransferase selectively methylates duplex DNA containing mispairs.** *Nucl. Acids Res.* **15**, 6,899-6,916.

Smith, S. S., Laayoun, A., Lingeman, R. G., Baker, D. J. & Riley, J. (1994). **Hypermethylation of telomere-like foldbacks at codon 12 of the human c-Ha-*ras* gene and the trinucleotide repeat of the *FMR-I* gene of fragile X.** *J. Mol. Biol.* **243**, 143-151.

Snell, R. G., MacMillan, J. C., Cheadle, J. P., Fenton, I., Lazarou, L. P., Davies, P., MacDonald, M. E., Gusella, J. F., Harper, P. S. & Shaw, D. J. (1993). **Relationship between trinucleotide repeat expansion and phenotypic variation in Huntington's disease.** *Nature Genet.* **4**, 393-397.

Snow, K., Tester, D. J., Krukeberg, K. E., Schaid, D. J. & Thibodeau, S. N. (1994). **Sequence analysis of the fragile X trinucleotide repeat: implications for the origin of the fragile X mutation.** *Hum. Mol. Genet.* **3**, 1,543-1,551.

Sobue, G., Doyu, M., Nakao, N., Shimada, N., Mitsuma, T., Maruyama, H., Kawakami, H. & Nakamura, S. (1996). **Homozygosity for Machado-Joseph disease gene enhances phenotypic severity.** *J. Neurol. Neurosurg. Psychiatry* **60**, 354-356.

Stallings, R. L. (1994). **Distribution of trinucleotide microsatellites in different categories of mammalian genomic sequence: implications for human genetic diseases.** *Genomics* **21**, 116-121.

Stark, G. R. & Wahl, G. M. (1984). **Gene amplification.** *Ann. Rev. Biochem.* **53**, 447-491.

Strand, M., Prolla, T. A., Liskay, R. M. & Petes, T. D. (1993). **Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair.** *Nature* **365**, 274-276.

Streisinger, G., Okada, Y., Emrich, J., Newton, J. & Tsugita, A. (1966). **Frameshift mutations and the genetic code.** *Cold Spring Harbor Symp. Quant. Biol.* **31**, 77-84.

Suen, I. S., Rhodes, J. N., Christy, M., McEwen, B., Gray, D. M. & Mitas, M. (1999). **Structural properties of Friedreich's ataxia d(GAA) repeats.** *Biochim. Biophys. Acta* **1444**, 14-24.

Sutcliffe, J. S., Nelson, D. L., Zhang, F., Pieretti, M., Caskey, C. T., Saxe, D. & Warren, S. T. (1992). **DNA methylation represses *FMR-1* transcription in fragile X syndrome.** *Hum. Mol. Genet.* **1**, 397-400.

Sutherland, G. R. (1985). **The enigma of the fragile X chromosome.** *Trends in Genetics* **1**, 108-112.

Sutherland, G. R. (1991a). **Chromosomal fragile sites.** *Genet. Anal. Tech. Appl.* **8**, 161-166.

Sutherland, G. R. & Baker, E. (1986). **Effects of nucleotides on expression of the folate sensitive fragile sites.** *Am. J. Med. Genet.* **23**, 409-417.

Sutherland, G. R., Baker, E. & Fratini, A. (1985). **Excess thymidine induces folate sensitive fragile sites.** *Am. J. Med. Genet.* **22**, 433-443.

Sutherland, G. R., Baker, E. & Richards, R. I. (1998). **Fragile sites still breaking.** *Trends Genet.* **14**, 501-506.

Sutherland, G. R., Haan, E. A., Kremer, E., Lynch, M., Pritchard, M., Yu, S. & Richards, R. I. (1991b). **Hereditary unstable DNA - a new explanation for some old genetic questions.** *Lancet* **338**, 289-292.

Sutherland, G. R. & Richards, R. I. (1992). **Anticipation legitimized: unstable DNA to the rescue.** *Am. J. Hum. Genet.* **51**, 7-9.

Sutherland, G. R. & Richards, R. I. (1995). **The molecular basis of fragile sites in human chromosomes.** *Curr. Opin. Genet. Dev.* **5**, 323-327.

Takeda, J., Ishii, S., Seino, Y., Imamoto, F. & Imura, H. (1989). **Negative regulation of human insulin gene expression by the 5'-flanking region in non-pancreatic cells.** *FEBS Lett.* **247**, 41-45.

Takiyama, Y., Igarashi, S., Rogaeva, E. A., Endo, K., Rogaev, E. I., Tanaka, H., Sherrington, R., Sanpei, K., Liang, Y., Saito, M., Tsuda, T., Takano, H., Ikeda, M., Lin, C., Chi, H., Kennedy, J. L., Lang, A. E., Wherret, J. R., Segawa, M., Nomura, Y., Yuasa, T., Weissenbach, J., Yoshida, M., Nishizawa, M., Kidd, K. K., Tsuji, S. & St. George-Hyslop, P. H. (1995). **Evidence for inter-generational instability in the CAG repeat in the MJD1 gene and for conserved haplotypes at flanking markers amongst Japanese and Caucasian subjects with Machado-Joseph disease.** *Hum. Mol. Genet.* **4**, 1,137-1,146.

Takiyama, Y., Sakoe, K., Soutome, M., Namekawa, M., Ogawa, T., Nakano, I., Igarashi, S., Oyake, M., Tanaka, H., Tsuji, S. & Nishizawa, M. (1997). **Single sperm analysis of the CAG repeats in the gene for Machado-Joseph disease (*MJD1*): evidence for non-Mendelian transmission of the *MJD1* gene and for the effect of the intragenic CGG/GGG polymorphism on the intergenerational instability.** *Hum. Mol. Genet.* **6**, 1,063-1,068.

Tautz, D. & Renz, M. (1984). **Simple sequences are ubiquitous repetitive components of eukaryotic genomes.** *Nucl. Acids Res.* **12**, 4,127-4,138.

Thaler, D. S., Stahl, M. M. & Stahl, F. W. (1987). **Tests of the double-strand break repair model for Red-mediated recombination of phage λ and plasmid λdv.** *Genetics* **116**, 501-511.

Thibodeau, S. N., Bren, G. & Schaid, D. (1993). **Microsatelite instability in familial cancer of the proximal colon.** *Science* **260**, 816-819.

Tilley, W. D., Marcelli, M., Wilson, J. D. & McPhaul, M. J. (1989). **Characterization and expression of a cDNA encoding the human androgen receptor.** *Proc. Natl Acad. Sci. U.S.A.* **86**, 327-331.

Tishkoff, D. X., Filosi, N., Gaida, G. M. & Kolodner, R. D. (1997). **A novel mutation avoidance mechanism dependent on S. cerevisiae *RAD27* is distinct from DNA mismatch repair.** *Cell* **88**, 253-263.

Tishkoff, S. A., Goldman, A., Calafell, F., Speed, W. C., Deinard, A. S., Bonne-Tamir, B., Kidd, J. R., Pakstis, A. J., Jenkins, T. & Kidd, K. K. (1998). **A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations.** *Am. J. Hum. Genet.* **62**, 1,389-1,402.

Trepicchio, W. L. & Krontiris, T. G. (1992). **Members of the *rel*/NF-κ B family of transcriptional regulatory proteins bind the *HRAS1* minisatellite DNA sequence.** *Nucl. Acids Res.* **20**, 2,427-2,434.

Trepicchio, W. L. & Krontiris, T. G. (1993). **IGH minisatellite suppression of USF-binding-site- and Eμ-mediated transcriptional activation of the adenovirus major late promoter.** *Nucl. Acids Res.* **21**, 977-985.

Trinh, T. Q. & Sinden, R. R. (1991). **Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli*.** *Nature* **352, 8th August**, 544-547.

Tsilfidis, C., MacKenzie, A. E., Mettler, G., Barcelo, J. & Korneluk, R. G. (1992). **Correlation between CTG trinucleotide repeat length and frequency of severe congenital myotonic dystrophy.** *Nature Genet.* **1**, 192-195.

Tsuji, S. (1997). **Molecular genetics of triplet repeats: Unstable expansion of triplet repeats as a new mechanism for neurodegenerative diseases.** *Internal Medicine* **36**, 3-8.

Usdin, K. (1998). **NGG-triplet repeats form similar intrastrand structures: implications for the triplet expansion diseases.** *Nucl. Acids Res.* **26**, 4,078-4,085.

Usdin, K. & Woodford, K. J. (1995). **CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis *in vitro*.** *Nucl. Acids Res.* **23**, 4,202-4,209.

van Dongen, M. J. P., Mooren, M. M. W., Willems, E. F. A., van der Marel, G. A., van Boom, J. H., Wijmenga, S. S. & Hilbers, C. W. (1997). **Structural features of the DNA hairpin d(ATCCTA-GTTA-TAGGAT): Formation of a G-A base pair in the loop.** *Nucl. Acids Res.* **25**, 1,537-1,547.

Varani, G. (1995). **Exceptionally stable nucleic acid hairpins.** *Ann. Rev. Biophys. Biomol. Struc.* **24**, 379-404.

Venczel, E. A. & Sen, D. (1993). **Parallel and antiparallel G-DNA structures from a complex telomeric sequence.** *Biochemistry* **32**, 6,220-6,228.

Vergnaud, G., Mariat, D., Apiou, F., Aurias, A., Lathrop, M. & Lauthier, V. (1991). **The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence.** *Genomics* **11**, 135-144.

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F. P., Eussen, B. E., Vanommen, G.-J. B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Caskey, C. T., Nelson, D. L., Oostra, B. A. & Warren, S. T. (1991). **Identification of a gene (*FMR-1*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome.** *Cell* **65**, 905-914.

Virtaneva, K., D'Amato, E., Miao, J., Koskiniemi, M., Norio, R., Avanzini, G., Franceschetti, S., Michelucci, R., Tassinari, C. A., Omer, S., Pennacchio, L. A., Myers, R. M., Dieguez-Lucena, J. L., Krahe, R., de la Chapelle, A. & Lehesjoki, A.-E. (1997). **Unstable**

minisatellite expansion causing recessively inherited myoclonus epilepsy, EPM1. *Nature Genet.* **15**, 393-396.

Waldman, A. S. & Liskay, R. M. (1987). **Differential effects of base-pair mismatch on intrachromosomal versus extrachromosomal recombination in mammalian cells.** *Proc. Natl Acad. Sci. U.S.A.* **84**, 5,340-5,344.

Warren, S. T. (1996). **The expanding world of trinucleotide repeats.** *Science* **271**, 1,374-1,375.

Warren, S. T., Zhang, F. P., Licameli, G. R. & Peters, J. F. (1987). **The fragile X site in somatic cell hybrids: an approach for molecular cloning of fragile sites.** *Science* **237**, 420-423.

Warrick, J. M., Paulson, H. L., Gray-Board, G. L., Bui, Q. T., Fischbeck, K. H., Pittman, R. N. & Bonini, N. M. (1998). **Expanded polyglutamine protein forms nuclear inclusions and causes neural degeneration in *Drosophila*.** *Cell* **93**, 939-949.

Weber, J. L. & Wong, C. (1993). **Mutation of human short tandem repeats.** *Hum. Mol. Genet.* **2**, 1,123-1,128.

Weitzmann, M. N., Woodford, K. J. & Usdin, K. (1996). **The development and use of a DNA polymerase arrest assay for the evaluation of parameters affecting intrastrand tetraplex formation.** *J. Biol. Chem.* **271**, 20,958-20,964.

Wells, R. D. (1996). **Molecular basis of genetic instability of triplet repeats.** *J. Biol. Chem.* **271**, 2,875-2,878.

Wells, R. D., Büchi, H., Kössel, H., Ohtsuka, E. & Khorana, H. G. (1967b). **Studies on polynucleotides. LXX. Synthetic deoxyribopolynucleotides as templates for the DNA polymerase of *Escherichia coli*: DNA-like polymers containing repeating tetranucleotide sequences.** *J. Mol. Biol.* **27**, 265-272.

Wells, R. D., Jacob, T. M., Narang, S. A. & Khorana, H. G. (1967a). **Studies on polynucleotides. LXIX. Synthetic deoxyribopolynucleotides as templates for the DNA polymerase of *Escherichia coli*: DNA-like polymers containing repeating trinucleotide sequences.** *J. Mol. Biol.* **27**, 237-263.

Wells, R. D., Parniewski, P., Pluciennik, A., Bacolla, A., Gellibolian, R. & Jaworski, A. (1998). **Small slipped register genetic instabilities in *Escherichia coli* in triplet repeat sequences associated with hereditary neurological diseases.** *J. Biol. Chem.* **273**, 19,532-19,541.

Willems, P. J. (1994). **Dynamic mutations hit double figures.** *Nature Genet.* **8**, 213-215.

Williams, N. G., Williams, L. D. & Shaw, B. R. (1989). **Dimers, trimers, and tetramers of cytosine with guanine.** *J. Am. Chem. Soc.* **111**, 7,205-7,209.

Williamson, J. R. (1994). **G-quartet structures in telomeric DNA.** *Annu. Rev. Biophys. Biomol. Struct.* **23**, 703-730.

Winter, R. M., Pembrey, M. E. & Davies, K. E. (1985). **Analysis of linkage data with factor IX corroborates the hypothesis that a premutation, followed by a recombination event, generates the full mutation in fragile X linked mental retardation.** *J. Med. Genet.* **22**, 397-397.

Woese, C. R., Winker, S. & Gutell, R. R. (1990). **Architecture of ribosomal RNA: Constraints on the sequence of "tetra-loops".** *Proc. Natl Acad. Sci. U.S.A.* **87**, 8,467-8,471.

Wöhrle, D., Kennerknecht, I., Wolf, M., Enders, H., Schwemmle, S. & Steinbach, P. (1995). **Heterogeneity of DM kinase repeat expansion in different fetal tissues and further expansion during cell proliferation *in vitro* : evidence for a causal involvement of methyl-directed DNA mismatch repair in triplet repeat stability.** *Hum. Mol. Genet.* **4**, 1,147-1,153.

Wolters, J. (1992). **The nature of preferred hairpin structures in 16S-like rRNA variable regions.** *Nucl. Acids Res.* **20**, 1,843-1,850.

Wong, Z., Wilson, V., Patel, I., Povey, S. & Jeffreys, A. J. (1987). **Characterization of a panel of highly variable minisatellites cloned from human DNA.** *Ann. Hum. Genet.* **51**, 269-688.

Woodford, K. J., Howell, R. M. & Usdin, K. (1994). **A novel K^+-dependent DNA synthesis arrest site in a commonly occurring sequence motif in eukaryotes.** *J. Biol. Chem.* **269**, 27,029-27,035.

Yamagata, H., Miki, T., Ogihara, T., Nakagawa, M., Higuchi, I., Osame, M., Shelbourne, P., Davies, J. & Johnson, K. (1992). **Expansion of unstable DNA region in Japanese myotonic dystrophy patients.** *Lancet* **339**, 692.

Yoshizawa, S., Kawai, G., Watanabe, K., Miura, K.-i. & Hirao, I. (1997). **GNA trinucleotide loop sequences producing extraordinarily stable DNA minihairpins.** *Biochemistry* **36**, 4,761-4,767.

Yu, A., Barron, M. D., Romero, R. M., Christy, M., Gold, B., Dai, J., Gray, D. M., Haworth, I. S. & Mitas, M. (1997b). **At physiological pH, d(CCG)$_{15}$ forms a hairpin containing protonated cytosines and a distorted helix.** *Biochemistry* **36**, 3,687-3,699.

Yu, A., Dill, J. & Mitas, M. (1995b). **The purine-rich trinucleotide repeat sequences d(CAG)$_{15}$ and d(GAC)$_{15}$ form hairpins.** *Nucl. Acids Res.* **23**, 4,055-4,057.

Yu, A., Dill, J., Wirth, S. S., Huang, G., Lee, V. H., Haworth, I. S. & Mitas, M. (1995a). **The trinucleotide repeat sequence d(GTC)$_{15}$ adopts a hairpin conformation.** *Nucl. Acids Res.* **23**, 2,706-2,714.

Yu, S., Mangelsdorf, M., Hewett, D., Hobson, L., Baker, E., Eyre, H. J., Lapsys, N., Le Paslier, D., Doggett, N. A., Sutherland, G. R. & Richards, R. I. (1997a). **Human chromosomal fragile site *FRA16B* is an amplified AT-rich minisatellite repeat.** *Cell* **88**, 367-374.

Yu, S., Mulley, J., Loesch, D., Turner, G., Donnelly, A., Gedeon, A., Hillen, D., Kremer, E., Lynch, M., Pritchard, M., Sutherland, G. R. & Richards, R. I. (1992). **Fragile-X syndrome: unique genetics of the heritable unstable element.** *Am. J. Hum. Genet.* **50**, 968-980.

Yu, S., Pritchard, M., Kremer, E., Lynch, M., Nancarrow, J., Baker, E., Holman, K., Mulley, J. C., Warren, S. T., Schlessinger, D., Sutherland, G. R. & Richards, R. I. (1991). **Fragile X genotype characterized by an unstable region of DNA.** *Science* **252**, 1,179-1,181.

Zerylnick, C., Torroni, A., Sherman, S. L. & Warren, S. T. (1995). **Normal variation at the myotonic dystrophy locus in global human populations.** *Am. J. Hum. Genet.* **56**, 123-130.

Zhao, Y. F., Cheng, W. J., Gibb, C. L. D., Gupta, G. & Kallenbach, N. R. (1996). **HMG Box proteins interact with multiple tandemly repeated (GCC)$_n$·(GGC)$_m$ DNA sequences.** *J. Biomol. Struc. Dyn.* **14**, 235-238.

Zheng, M., Huang, X., Smith, G. K., Yang, X. & Gao, X. (1996). **Genetically unstable CXG repeats are structurally dynamic and have a high propensity for folding. An NMR and UV spectroscopic study.** *J. Mol. Biol.* **264**, 323-336.

Zhu, L., Chou, S.-H., Xu, J. & Reid, B. R. (1995). **Structure of a single-cytidine hairpin loop formed by the DNA triplet GCA.** *Nature Struct. Biol.* **2**, 1,012-1,017.

Zhu, Q.-S., Heisterkamp, N. & Groffen, J. (1990). **Unique organization of the human BCR gene promoter.** *Nucl. Acids Res.* **18**, 7,119-7,125.

Zhuchenko, O., Bailey, J., Bonnen, P., Ashizawa, T., Stockton, D. W., Amos, C., Dobyns, W. B., Subramony, S. H., Zoghbi, H. Y. & Lee, C. C. (1997). **Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the α_{1A}-voltage-dependent calcium channel.** *Nature Genet.* **15**, 62-69.

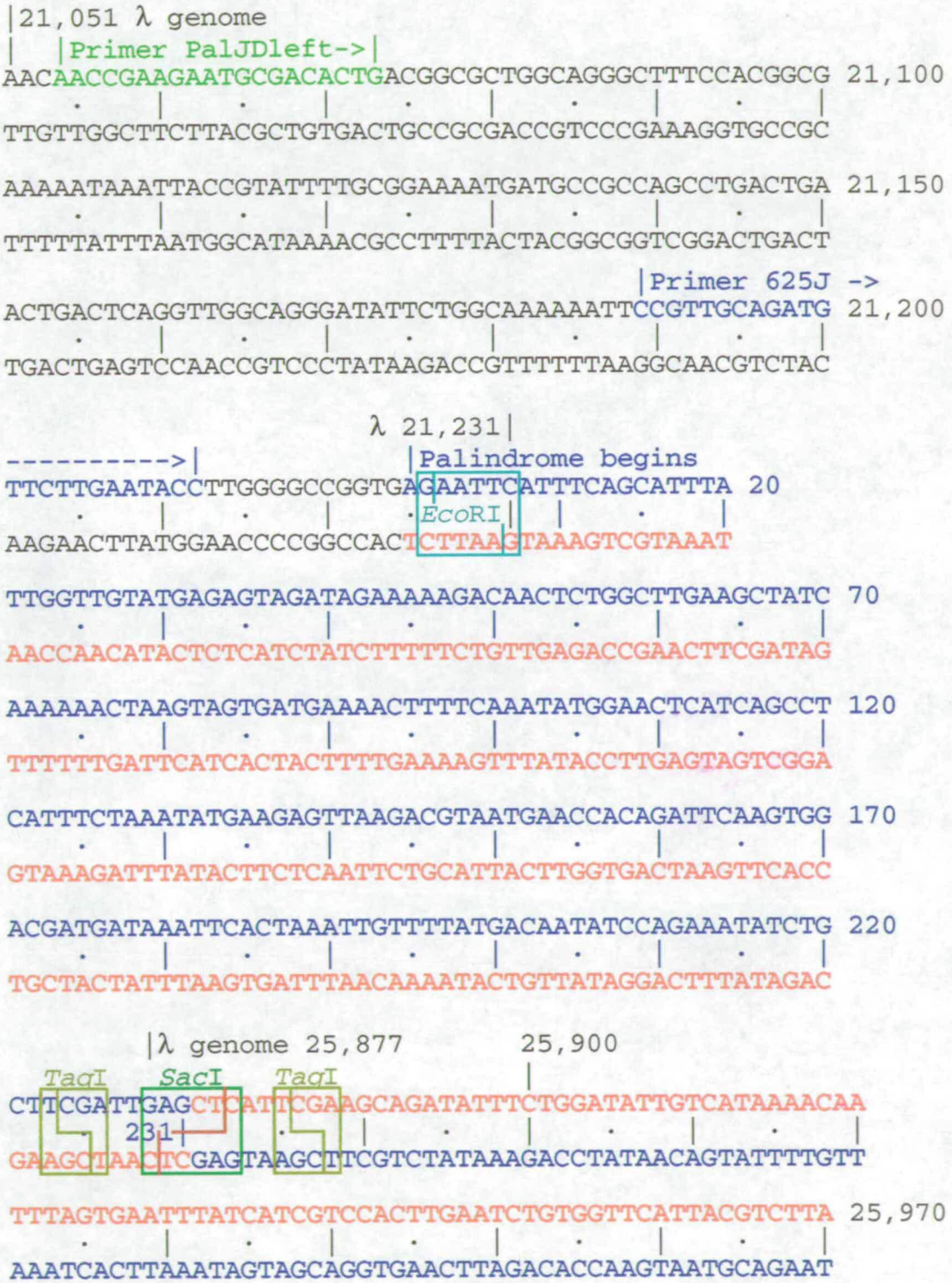
Zischler, H., Kammerbauer, C., Studer, R., Grzeschik, K.-H. & Epplen, J. T. (1992). **Dissecting (CAC)$_5$/(GTG)$_5$ multilocus fingerprints from man into individual locus-specific, hypervariable components.** *Genomics* **13**, 983-990.

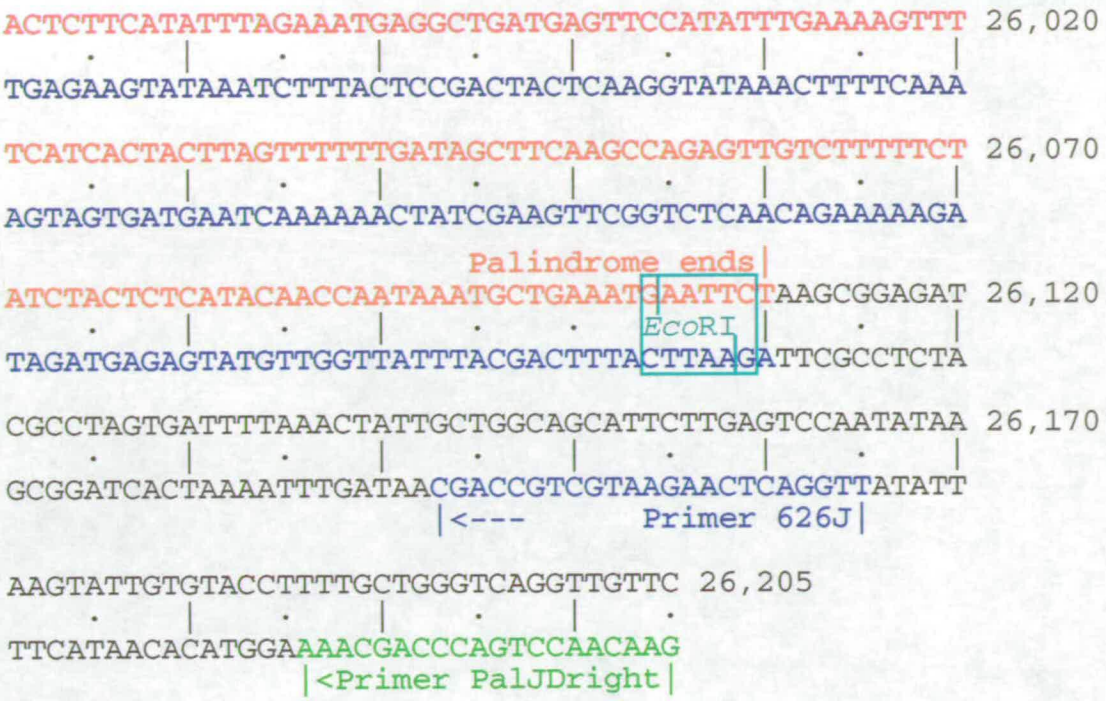
Zoghbi, H. Y. (1997). **Molecular genetics and neurobiology of neurodegenerative and neurodevelopmental disorders.** *Pediatric Research* **41**, 722-726.

Zühlke, C., Riess, O., Bockel, B., Lange, H. & Thies, U. (1993). **Mitotic stability and meiotic variability of the (CAG)$_n$ repeat in the Huntington disease gene.** *Hum. Mol. Genet.* **2**, 2,063-2,067.

Appendix 1

The palindrome of DRL167 and surrounding sequence





Nucleotide sequence of the 462 bp palindrome in DRL167 and flanking sequence showing primers used between the outside of the palindrome and a primer complementary to a sequence ligated into the centre (see Chapters 2 and 7). The λ sequence between the *Eco*RI site at 21,226 - 21,231 and the *Sac*I site at 25,877 - 25,882 is deleted. The first half of the palindrome is an inverted repeat of the λ sequence between the *Sac*I site and the next *Eco*RI site at 26,104 - 26,109 +, by chance, 1 further base-pair at each end. The palindrome sequence is coloured in blue and red to emphasize the inverted repeat nature. 5'→3' the blue sequences are identical; ditto the red. The flanking λ sequence is in black. λ sequence is numbered in black and the left half of the palindrome is numbered in blue, *i.e.* 1 - 231 making the total palindrome 231 × 2 = 462.

Appendix 2

e.mail to Dr. Dinshaw J. Patel

From: Self <BIO-SRV0/JDARLOW>
To: pateld@mskcc.org
Subject: Some questions on Kettani et al. (1995)
Date: Tue, 11 Mar 1997 22:02:17

Dear Dr. Patel,

I've been trying to make sense of all the papers which come to different conclusions on the natures of secondary structures formed by the single strands of d(CGG).d(CCG) repeats and I'm very interested by your paper Kettani et al. (1995) but there are some questions to which I haven't found the answers and I'd be pleased if you could enlighten me.

1. The bonding between the two C.G pairs in C.G.C.G tetrads: Löwdin (1964) and Kubitschek & Henderson (1966) proposed hydrogen bonds from the N4 of cytosine (incorrectly labelled 6 by K & H) to the O6 of the guanine of the other C.G pair and this was the pairing that Mitas et al. (Biochemistry, 1995) obtained from computer modelling. O'Brien (1967), whom you quoted, found by X-ray crystallography of 9-methyl guanine and 1-methylcytosine that bonding was with the N7s of the guanines. McGavin (1971), whom you also quoted, drew an arrangement half-way between the two bonding schemes but commented that the NH---N distance was probably too long for the O'Brien scheme and the angle between NH and NO was probably too large for the alternative bonding scheme and from his discussion he seemed to consider that the bonding had to be one or the other. However, you have stated that the cytosine amino protons donate bifurcated hydrogen bonds to O6 and N7 of guanine. What is your basis for this? Is it just a computer model or can you really discern it from the NMR data?

2. I am very interested by your statement in the first section of the discussion that in contrast to the spectacular proton spectrum of GCGGTTTGCGG the proton spectrum of GGCGTTTGGCG was of poor quality with broad resonances and multiple conformations. I cannot find any mention of this latter molecule in the results and there is no mention of its synthesis in the materials and methods section.

Where are the results for GGCGTTTGGCG; is there another paper that I haven't seen?

3. It seems to me that one of the reasons that different teams found different structures for the G-rich strand [(GGC)n] is that some people used too low a temperature or waited too short a time for quadruplex formation. How long did you redisolve your lyophilized purified oligonucleotides and at what temperature?

Thanks very much,

John Darlow
Institute of Cell and Molecular Biology,
University of Edinburgh, Scotland.

The Effects of Trinucleotide Repeats Found in Human Inherited Disorders on Palindrome Inviability in *Escherichia coli* Suggest Hairpin Folding Preferences *In Vivo*

John M. Darlow and David R. F. Leach

Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh EH9 3JR, United Kingdom

Manuscript received July 13, 1995
Accepted for publication July 29, 1995

ABSTRACT

Unusual DNA secondary structures have been implicated in the expansion of trinucleotide repeat tracts that are associated with several human inherited disorders. We present evidence consistent with the folding of these trinucleotide repeats into hairpin loops at the center of a long DNA palindrome *in vivo*. Our assay utilizes a palindrome in bacteriophage λ, the center of which determines its ability to inhibit plaque formation in a manner that is consistent with folding into a hairpin or cruciform structure. We show that central inserts of even numbers of d(CAG)·d(CTG) repeats inhibit plaque formation more than do odd numbers. Both $d(CAG)_2 \cdot d(CTG)_2$ and $d(CGG)_2 \cdot d(CCG)_2$ central sequences behave like DNA sequences known to form two-base loops *in vitro*, suggesting that they may also form compact and stable loops. By contrast, repeats of d(GAC)·d(GTC) do not show any evidence consistent with unusual loop stability. These results agree with *in vitro* evidence that the unstable repeats can form hairpin secondary structures and suggest a favored position of folding. We discuss the potential roles of secondary structures, DNA replication and recombination in models of repeat tract expansion.

IN the past four years, ten human inherited disorders and/or fragile chromosome sites have been shown to be caused by amplification of trinucleotide repeats (dynamic mutation) (WILLEMS 1994; RITCHIE *et al.* 1994). Of the ten possible trinucleotide repeats (when strand and frame are ignored) only two occur at these ten loci, and both are of the form d(C*X*G)·d(C*X*′G), where *X* and *X*′ are complementary bases. At all these loci there is polymorphism of the number of repeat units in normal individuals; the chance of expansion is related to the length of uninterrupted repeats, and beyond a certain threshold, symptoms and/or fragile site expression appear. Such alleles are said to have the full mutation and are highly polymorphic and unstable. The d(CGG)·d(CCG) repeats are all associated with fragile sites (two of them associated with mental retardation) and full mutations are massive with hundreds or even thousands of repeats. The d(CAG)·d(CTG) repeat disorders are all progressive neuromuscular disorders, and an increase in the number of repeats in successive generations is associated, to varying degrees, with increase in severity and decrease in onset age (anticipation).

Two different classes of mechanism have been proposed to account for dynamic mutation: strand-slippage and recombination. Prototype models of these mechanisms are shown in Figure 1. The concept, variously known as "strand-slippage", "replication-slippage", and "slipped-strand mispairing", was first proposed by STREISINGER *et al.* (1966). In this the growing tip of a nascent DNA strand is displaced between *direct* repeats on the parent strand, resulting in insertion or deletion of bases on the new strand depending upon the direction of the displacement. LEVINSON and GUTMAN (1987) proposed that some repeat sequences, through being quasi-palindromic, might expand by a mechanism involving self-complementarity, *i.e.*, if direct repeats are partially palindromic, intrastrand pairing may occur and this may aid strand-slippage. This mechanism has been put forward specifically for strand-slippage during replication of trinucleotide repeats (SINDEN and WELLS 1992). It has been proposed that the increased frequency of expansion observed for long tracts of repeats may relate to their length being in excess of that observed for Okazaki fragments (RICHARDS and SUTHERLAND 1994). These authors suggested that an Okazaki fragment composed exclusively of trinucleotide repeats might be able to slide on its template and thereby allow its expansion. The mechanism of sliding was not discussed, but we suggest that this apparent sliding may be facilitated by intrastrand pairing of the Okazaki fragment and its template, as shown in Figure 1. The alternative of concerted melting of a whole Okazaki fragment and its reannealing in a new position seems implausible. A recombinational mechanism for repeat instability may also involve DNA secondary structure. JANSEN *et al.* (1994) suggest that an unusual DNA secondary structure causes DNA double-strand cleavage in *trans* on a sister chromatid to initiate recombination. We propose that cleavage in *cis* at the site of a secondary

Corresponding author: Dr. David R. F. Leach, Institute of Cell and Molecular Biology, Darwin Bldg., King's Buildings, Mayfield Rd., Edinburgh EH9 3JR, United Kingdom. E-mail: d.leach@ed.ac.uk

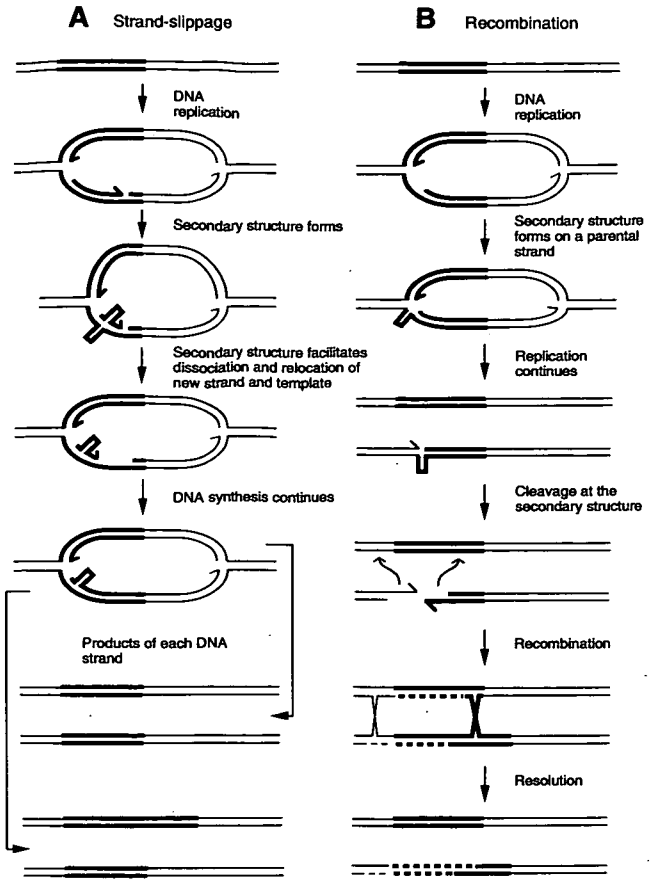


FIGURE 1.—Two classes of model for dynamic mutation nd the potential involvement of unusual DNA secondary tructures in each. The tract of trinucleotide repeats is repreented by thick lines and normal DNA by thin lines. The econdary structure is represented by a hairpin but may be nore complex. (A) Replication-slippage can account for amlification if the newly synthesized strand folds back on itself nd then replication recopies the section of template preiously used. Slippage is represented on the lagging-strand f the replication fork since several studies suggest that for airpins formed by palindromic DNA, slippage occurs more requently there than on the leading strand (LEACH 1994). B) Recombination is known to be stimulated at sites of doule-strand breaks. In a tract of trinucleotide repeats, the formation of a secondary structure may lead to cleavage of one sister chromatid and its repair by recombination. The broken arm may recombine at many sites within the repeated array to generate new alleles of variable length. Resolution is shown by disengagement of strands as proposed in several double-strand break repair models (RESNICK 1976; NASMYTH 1982; THALER *et al.* 1987; HASTINGS 1988). If recombination with a homologous chromosome is possible, this is necessary to explain the lack of crossing over associated with repeat expansion (IMBERT *et al.* 1993; JEFFREYS *et al.* 1994; KUNST and WARREN 1994). Resolution by cleavage of the Holliday junctions is a viable alternative if recombination is primarily or exclusively between sister chromatids.

structure formed on the lagging strand of a replication fork is more plausible (see Figure 1).

We have used a bacteriophage λ derivative containing a long palindrome to study behavior of trinucleotide repeats *in vivo*. In double-stranded DNA palindromes, *i.e.*, inverted repeats, opposite halves of the palindrome on the *same* strand are complementary to one another and may anneal to form a hairpin or a cruciform if both strands so pair. Long DNA palindromes are not recovered in DNA libraries when they are introduced into bacteria. Either they are wholly or partially deleted (instability) or they cause failure of replication of the vector (inviability) (LEACH 1994). The threshold for this inviability is ~150–200 bp total length of the palindrome. Further investigations have led to the discovery that mutations in *Escherichia coli* genes *sbcC* and *sbcD* allow the propagation of bacteriophage λ derivatives with long palindromes (CHALKER *et al.* 1988; GIBSON *et al.* 1992). However, in *sbcC* mutant hosts inviability is not totally overcome and the plaque size of palindrome-containing phage is acutely sensitive to the central sequence of the palindrome (DAVISON and LEACH 1994a). DAVISON and LEACH (1994a) showed that central sequences predicted to stabilize DNA hairpins reduce plaque size. In positions outside the central two base-pairs of a perfect palindrome, C and G produced smaller plaques than A and T. This is the reverse of what would be expected if melting were the rate-limiting step, as it appears to be *in vitro*. The observed effect diminishes with distance from the center, suggesting that formation of the first few intrastrand base pairs ("protocruciform" formation) is the rate limiting step *in vivo*. Also, sequences known to adopt two-base loops *in vitro* generate smaller plaques than sequences known to adopt four-base loops *in vitro* (DAVISON and LEACH 1994b). These studies have shown that there is no correlation between predicted central melting and plaque size, but that hairpin-loop stability correlates inversely with plaque size. We have concluded, therefore, that the measurement of plaque areas is a reliable assay for the stability of DNA hairpin loops (DAVISON and LEACH 1994a,b).

The model that d(C*X*G)·d(C*X*′G) repeats may form a pseudo-hairpin held together by C·G pairing proposed by SINDEN and WELLS (1992) has recently been supported by *in vitro* gel electrophoretic and NMR analysis of oligonucleotide sequences (SMITH *et al.* 1994; CHEN *et al.* 1995; GACY *et al.* 1995; MITAS *et al.* 1995). These pseudo-hairpins have the potential to exist in two possible forms comprising odd or even numbers of repeat units that differ only in the nature of the loop formed at the apex of the hairpin (LEACH 1994) (see Figure 2). To test whether one or other of these loops might be particularly stable *in vivo*, we have compared the behavior of λ phages containing different numbers of d(CAG)·d(CTG) repeats inserted into the center of a long palindrome. We have also looked at the behavior of odd and even numbers of d(GAC)·d(GTC) and d(CGG)·d(CCG) repeats. Relative to d(CAG)·d(CTG), the d(GAC)·d(GTC) repeats consist of a change in position of C and G bases that generates a different trinucleotide (not simply a circular permutation of the

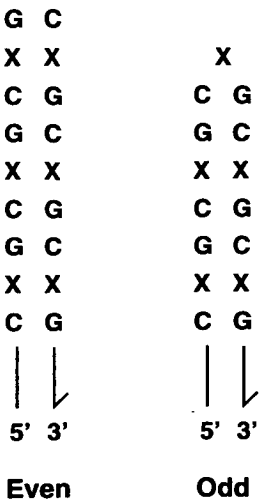
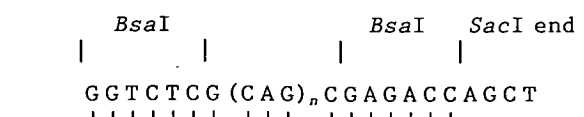


FIGURE 2.—Even and odd forms of pseudo-hairpins stabilized by C·G base pairing in the repeated sequence $d(CXG)_n$ where *X* represents one of the bases A, T, C or G. If *X* is T or C, some non-Watson-Crick hydrogen bonding between identical bases may be possible and stacking within the helix is likely. If *X* is A or G, it may be more difficult to accommodate the bulky purines opposite each other in the helix. Note that these forms are only distinguished by the loops at the apex of the pseudo-hairpins.

test sequence) but that still has the same potential for C·G pairing within a pseudo-hairpin. Furthermore, only runs of five or fewer repeats of this trinucleotide have been detected in human DNA sequences (GACY *et al.* 1995). We show that increasing numbers of d(GAC)·d(GTC) repeats at the center progressively increase plaque size, suggesting that they do not form unusually stable hairpin loops. In contrast, even and odd numbers of d(CAG)·d(CTG) repeats behave differently. Even numbers of repeats give rise to smaller plaques, suggesting that a favored position of folding exists between the repeats. The sequence $d(CAG)_2 \cdot d(CTG)_2$ behaves as though it is able to form a particularly stable loop, and we have shown that the plaque size is the same as that resulting from a sequence known to form a two-base loop *in vitro*. We also find that $d(CGG)_2 \cdot d(CCG)_2$ behaves as though it can form a compact stable loop.

MATERIALS AND METHODS

Bacteriophage and oligonucleotides: A bacteriophage λ construct, DRL167 (*pal, spi6, cI857, χCI53*) (DAVISON and LEACH 1994a), was used. This contains a 462-bp perfect palindrome with a *SacI* site at the center. Double-stranded DNA ligated into this site was made by annealing complementary oligonucleotides to make inserts of the following form:



Only the $d(CAG)_n \cdot d(CTG)_n$ trinucleotide is shown. The others used were $d(GAC)_n \cdot d(GTC)_n$ and $d(CGG)_n \cdot d(CCG)_n$. The use of an asymmetric indicator restriction enzyme site (*BsaI*) in opposite orientations on either side of the center ensures that there is no competing eccentric folding site for the surrounding palindrome arms, while continuing the palindromic sequence right up to the trinucleotides. The *SacI* site is destroyed by the insert.

Plaque size assays: These were carried out as described (DAVISON and LEACH 1994b) with the following exceptions and additions. The agar contained not casitone but 10 g BBL Select Trypticase Peptone (Becton Dickinson) per liter and 10 mM Tris·HCl (pH 7.5). The plates were poured to a volume of exactly 40 ml and 2.5 ml of top agar was used. The plates were allowed to dry in stacks of >20 and after 3 days drying were not randomly ordered but dealt like cards into the required number of piles, with five plates per pile and one pile for each phage isolate being assayed at the time. Usually plaque areas were measured on five plates per phage isolate, but two isolates of phage were quantified for several of the 19 different inserts and paired sets of results were always similar. As the plaque size assays were not all done at the same time, the same control was always run, another bacteriophage λ derivative (DRL176) containing a 476-bp palindrome, for plaque size comparison.

RESULTS

Oligonucleotides containing one to five copies of d(CAG)·d(CTG) were inserted into the center of a long palindrome in bacteriophage λ and their effect on plaque formation on an *E. coli sbcC* mutant host was observed. Cumulative frequency curves of plaque areas are shown in Figure 3 that demonstrate that even numbers of d(CAG)·d(CTG) repeats produce smaller plaques than do odd numbers and that the sequence $d(CAG)_2 \cdot d(CTG)_2$ gives very small plaques. When median plaque area is plotted against the number of repeat units, the alternation of plaque size for odd and even repeat numbers is clear (Figure 4A). This figure also shows that by contrast the d(GAC)·d(GTC) repeats show no evidence of odd-even alternation. A continuous increase in plaque size is observed as the number of these repeats is raised from one to five.

The fact that phage with a $d(CAG)_2 \cdot d(CTG)_2$ central sequence formed very small plaques suggested that this sequence might favor hairpin-loop formation. Were this a general feature of $d(CXG)_2 \cdot d(CX'G)_2$, then $d(CGG)_2 \cdot d(CCG)_2$ should also be a good folding sequence for hairpin formation. We therefore inserted one to five copies of d(CGG)·d(CCG) into the center of the same long palindrome in phage λ and observed the effect of these insertions on plaque formation. As with the d(CAG)·d(CTG) repeats, a central insert of $d(CGG)_2 \cdot d(CCG)_2$ resulted in the formation of small plaques suggesting that a stable loop was formed. However, the alternation of odd and even repeat numbers on plaque size did not extend to three, four and five repeats (Figure 4B). These results may suggest that other folding possibilities are available to $d(CGG)_n \cdot d(CCG)_n$ at $n \geq 4$ (see DISCUSSION).

The observation that central insertions of $d(CAG)_2 \cdot$

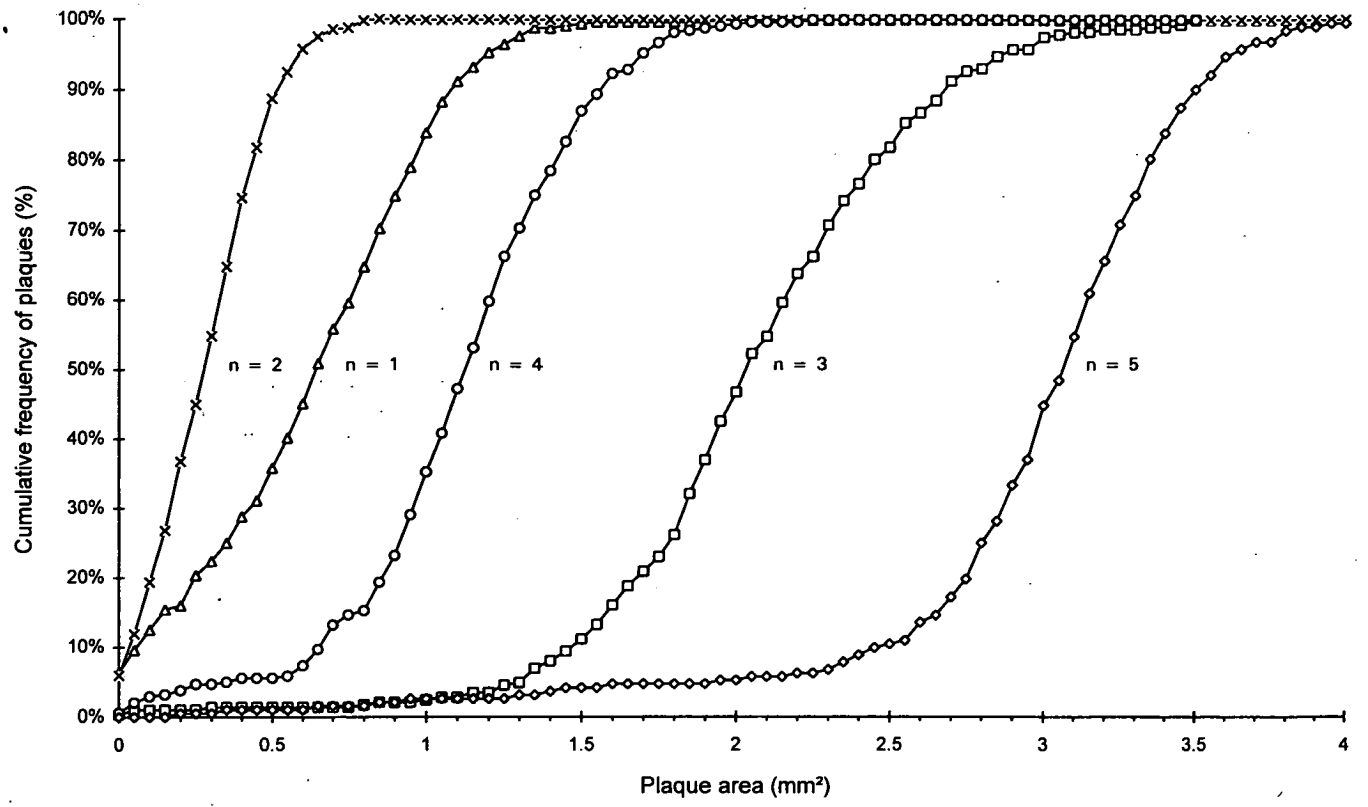


FIGURE 3.—Cumulative frequency (percentage) curves of plaque areas of phage with long palindromes, containing the respective numbers (n) of d(CAG)·d(CTG) trinucleotides in their centers, on *sbcC* mutant *E. coli*. Curves for one isolate of phage with each $d(CAG)_n \cdot d(CTG)_n$-containing insert are shown.

$d(CTG)_2$ and $d(CGG)_2 \cdot d(CCG)_2$ were responsible for the formation of very small plaques suggested that these sequences might be able to form tight loops. We have therefore compared the effects on plaque size of a $d(CAG)_2 \cdot d(CTG)_2$ central sequence with central sequences believed to form two- and four-base loops *in vitro* (HILBERS *et al.* 1994) and *in vivo* (DAVISON and LEACH 1994b). For this comparison, we reconstructed two of the central insertions used by DAVISON and LEACH (1994b) in the same sequence context that we have used in the present study and compared plaque sizes with the phage previously studied, and phage with central a $d(CAG)_2 \cdot d(CTG)_2$ insert. We also measured the plaque size of a phage with a central $d(GAC)_2 \cdot d(GTC)_2$ sequence. The results of this comparison are shown in Figure 5, where it can be seen that the $d(CAG)_2 \cdot d(CTG)_2$ central sequence confers a plaque size that is consistent with the formation of a two-base loop whereas the $d(GAC)_2 \cdot d(GTC)_2$ central sequence gives larger plaques even than the phage containing the sequence known to form four-base loops. This may indicate that this sequence prefers to exist in a loop with six unpaired bases.

For all of the central inserts studied, it was necessary to make a choice for the two base pairs flanking the $d(CXG)_n \cdot d(CX'G)_n$ sequence. These were chosen to generate the sequence $dG(CXG)_nC \cdot dG(CX'G)_nC$ because a 5′G and 3′C would be present in a long array flanking a trinucleotide of this sequence. However, the sequence $d(GAC)_n \cdot d(GTC)_n$ has a reversal of the G and C bases at the 5′ and 3′ ends of the repeat and a flanking 5′G and 3′C would not be the bases found adjacent to the trinucleotide in a repeated array. We have therefore considered the possibility that the small plaque phenotype conferred by $d(CAG)_2 \cdot d(CTG)_2$ might be due to the nature of the flanking bases. We have compared the plaque sizes of phages with the following four central sequences: $dG(CAG)_2C \cdot dG(CTG)_2C$, $dG(GAC)_2C \cdot dG(GTC)_2C$, $dC(CAG)_2G \cdot dC(CTG)_2G$ and $dC(GAC)_2G \cdot dC(GTC)_2G$ (see Figure 5). These results argue that the orientation of the flanking bases does not greatly influence the plaque size.

DISCUSSION

Although many different repeated sequences exist in mammalian genomes, to date only the two possible trinucleotide repeats of the form $d(CXG)_n \cdot d(CX'G)_n$ have been found associated with inherited disease. Both of these repeats have the potential to form pseudo-hairpins stabilized by C·G base-pairing, and we have suggested that they can adopt either of the two forms shown in Figure 2 (LEACH 1994). One folds between d(CXG) units and contains an even number of repeats

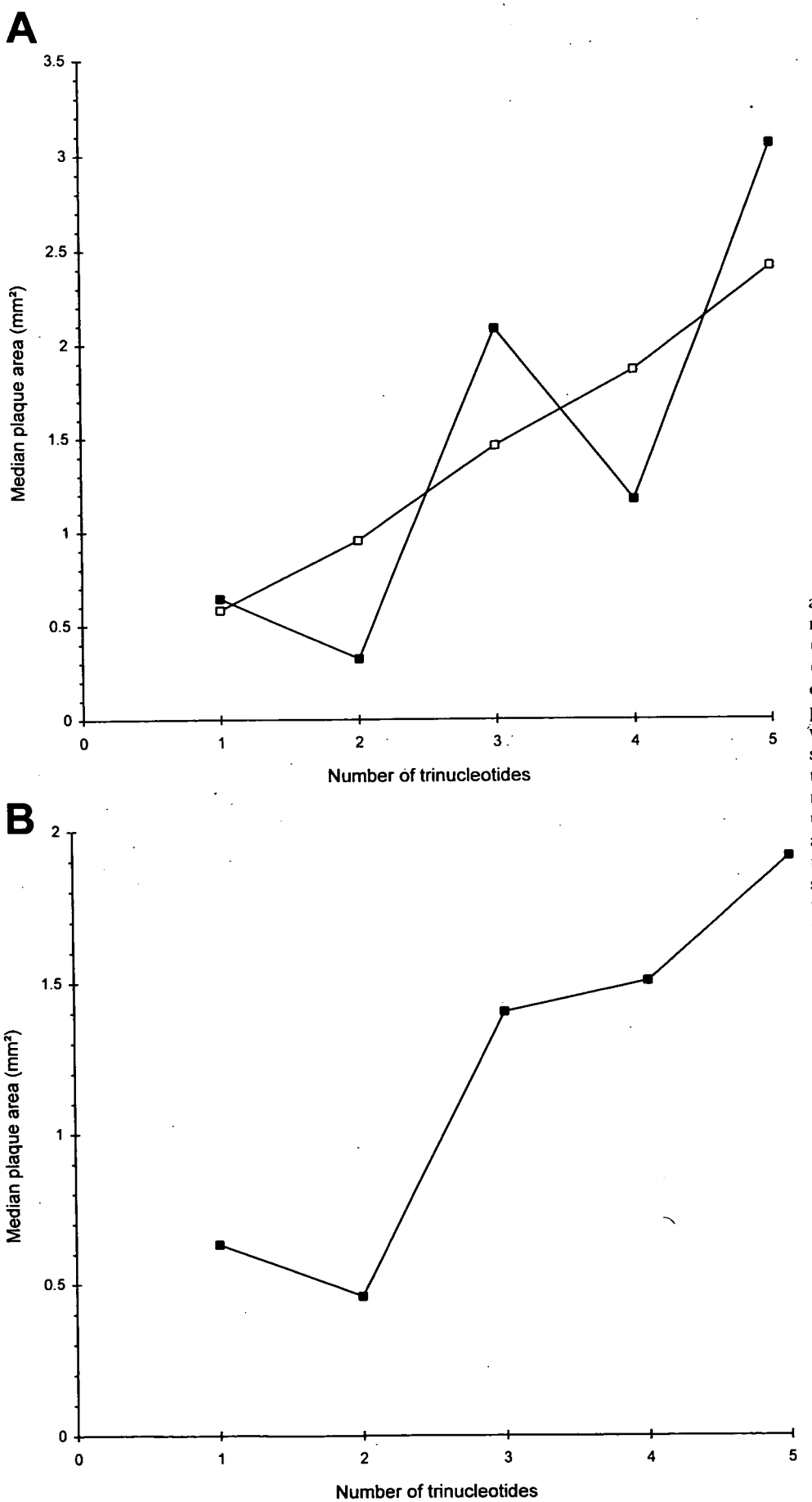


FIGURE 4.—Median plaque area plotted against number of trinucleotides. (A) $d(CAG)_n \cdot d(CTG)_n$ (■) and $d(GAC)_n \cdot d(GTC)_n$ (□). (B) $d(CGG)_n \cdot d(CCG)_n$. Whenever more than one isolate of a phage was assayed, the median was calculated from all plates measured. The sizes of the plaques in the $d(CGG)_n \cdot d(CCG)_n$ series may not be directly comparable with those of the $d(CAG)_n \cdot d(CTG)_n$ and $d(GAC)_n \cdot d(GTC)_n$ series as the assays were not done at the same time. For this reason no attempt has been made to equalize the scales of A and B.

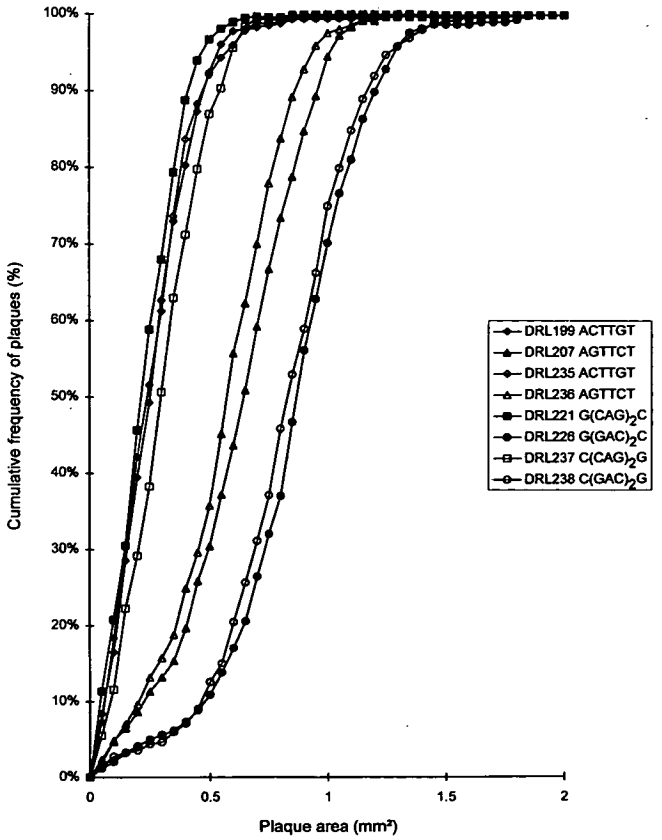


FIGURE 5.—Cumulative frequency curves of plaque areas of phage with central sequences known to form a two-base loop *in vitro* (DRL199) and a four-base loop *in vitro* (DRL207) and six other phage for comparison (discussed in text). DRL235 contains the same central six base pairs as DRL199 but within the context (insert) used in this study; DRL236 likewise corresponds to DRL207. All the assays on this graph were performed at the same time but at a different time to the previous assays, hence some difference in the median sizes of the DRL221 [$(CAG)_2$] and DRL 226 [$(GAC)_2$] here from those in Figure 4A.

and the other folds with an apical trinucleotide and has an odd number of repeats. If these pseudo-hairpin structures are prone to form in $d(CXG)_n \cdot d(CX'G)_n$ sequences, we predicted that one form would be more stable than the other and that this would be determined primarily by the stability of the loop at the apex of the hairpin. Furthermore, if $d(CXG)_n \cdot d(CX'G)_n$ sequences are particularly prone to form an unusual secondary structure, $d(GXC)_n \cdot d(GX'C)_n$ sequences that are not known to be prone to dynamic mutation might not favor secondary structure formation. Inside long tracts $d(CXG)_n \cdot d(CX'G)_n$ and $d(GXC)_n \cdot d(GX'C)_n$ are indistinguishable if X and X' are G or C but can be distinguished if they are A or T. We therefore set out to compare the *in vivo* properties of odd and even numbers of d(CAG) · d(CTG), d(GAC) · d(GTC) and d(CGG) · d(CCG) repeats at the center of a long palindrome in bacteriophage λ where previous studies have revealed an inverse correlation between hairpin-loop stability and plaque size.

It is possible that the behavior of d(CAG) · d(CTG) repeats may be determined by their potential to form pseudo-hairpins stabilized by C · G base-pairing. Both $d(CAG)_n$ and $d(CTG)_n$ single-strands form stable unimolecular pseudo-hairpin secondary structures that have been analyzed by gel electrophoresis and NMR (GACY *et al.* 1995; MITAS *et al.* 1995). These studies conclude that the pseudo-hairpins formed are stabilized by C · G base-pairing and, in the case of the $d(CTG)_n$ strand, that T · T base pairs are formed. Our *in vivo* experiments reveal that even repeat numbers of d(CAG) · d(CTG) at the center of a long palindrome produce smaller plaques than do odd numbers. This suggests that a favored position of folding may exist between pairs of these trinucleotides to generate an even-membered hairpin-loop that can now be studied *in vitro* by structural and thermodynamic methods. The sequence $d(CAG)_2 \cdot d(CTG)_2$ gives very small plaques, suggesting that it may fold into an unusually stable hairpin-loop. With higher numbers of d(CAG) · d(CTG) trinucleotides, there will be competing $d(CAG)_2 \cdot d(CTG)_2$ pairs on either side of the center, which may explain why $d(CAG)_4 \cdot d(CTG)_4$ plaques are larger than those of $d(CAG)_2 \cdot d(CTG)_2$. For odd numbers there are eccentric $d(CAG)_2 \cdot d(CTG)_2$ sites but no central site.

The behavior of $d(CGG)_n \cdot d(CCG)_n$ repeats may be more complex, both *in vitro* and *in vivo*. *In vitro* data suggest that the two strands of the $d(CGG)_n \cdot d(CCG)_n$ repeats prefer to adopt different conformations that involve folding in different frames. It has been shown that the C-rich single-strand can form a pseudo-hairpin stabilized by C · G base-pairing and containing C · C mispairs (SMITH *et al.* 1994; CHEN *et al.* 1995), while the G-rich strand forms hairpins also stabilized by C · G base-pairing but containing G · G mispairs. The C-rich strand prefers to pair as d(CCG) · d(CCG) (the frame used in our phage), but the G-rich strand prefers to pair as d(GGC) · d(GGC) (CHEN *et al.* 1995). Furthermore *in vitro* evidence consistent with the formation of a tetraplex structure by pairing of two $d(CGG)_n$ single-strand hairpins has also been obtained (FRY and LOEB 1994; SMITH *et al.* 1994) consistent with stabilization by G · G base-pairing (SINDEN and WELLS 1992) in the third possible frame of pairing. The fact that in our experiments $d(CGG)_n \cdot d(CCG)_n$ shows odd-even alternation at $n = 1$, $n = 2$ and $n = 3$ but not at $n = 3$, $n = 4$ and $n = 5$ may be explained by competition between alternative nucleation structures available to this sequence at $n \geq 4$. Any folding in an alternative frame is predicted to interfere with the pairing of the palindrome arms. By contrast, in a long array of trinucleotides the folding of the two strands in different frames is not problematic since each strand can continue to fold in its own frame until it reaches the end of the repeat tract.

In our system, no odd-even alternation of plaque size is observed for d(GAC) · d(GTC) repeats. Whether this relates to the observation that this sequence is not

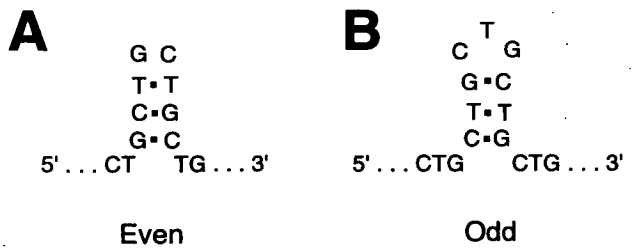


FIGURE 6.—Alternative loops that may be formed by one strand of $d(CTG)_n$ in hairpins stabilized by C·G base-pairing. The strands containing thymines are shown because the pyrimidine-containing strands are likely to form the most stable hairpin loops. Similar interactions may also occur in $d(CCG)_n$ strands. The figure is drawn to indicate the first three potential intrastrand base pairs. (A) An even number of d(CTG) repeats generates a structure with an axis of folding between the central two trinucleotides. The loop consists of the four-base sequence d(TGCT) closed by a C·G base pair. The C is on the 5′ side of the loop and the G on the 3′ side. This arrangement of 5′-pyrimidine 3′-purine closing the loop is particularly favorable (DAVISON and LEACH 1994b; HILBERS *et al.* 1994). The thymines are likely to be easily accommodated within the helix to form both stacking interactions with the loop-closing base pair and hydrogen-bonding interactions with each other as occurs in the wobble base pair observed between thymines 1 and 4 in a d(TTTT) loop (HILBERS *et al.* 1994). The structure may therefore share characteristics with two-base loops and is drawn with a T·T wobble base pair between bases 1 and 4 of the central loop. (B) An odd number of d(CTG) repeats generates an axis of folding that bisects the central trinucleotide. If a three-base loop forms, it will be closed by a G·C base pair. This base pair, with a guanine on the 5′ side of the loop and a cytosine on the 3′ side, is the less favored polarity for a loop-closing base pair (DAVISON and LEACH 1994; HILBERS *et al.* 1994).

found in expanded arrays ($n > 5$) has yet to be determined.

The tiny plaques formed by phage with $d(CAG)_2 \cdot d(CTG)_2$ and $d(CGG)_2 \cdot d(CCG)_2$ central sequences suggest that these arrangements of bases may form unusually stable hairpin loops. The hairpin loops formed by these $d(CXG)_2 \cdot d(CX'G)_2$ sequences consist of a 5′C and 3′G closing a four-membered loop (Figure 6A). This arrangement of loop-closing bases, which does not occur in the even fold of the sequence, $d(GAC)_2 \cdot d(GTC)_2$, has been shown to favor loop stability *in vitro* (HILBERS *et al.* 1994) and small plaques *in vivo* (DAVISON and LEACH 1994b). On the pyrimidine-containing strands, $d(CTG)_n$ or $d(CCG)_n$, the pair of thymines or cytosines one base removed from the center may be able to stack within the loop to leave only two effectively unpaired bases as proposed in Figure 6A. To determine whether the plaque-size data were consistent with the formation of two-base loops, we have compared the behavior of phage with central inserts of $d(CAG)_2 \cdot d(CTG)_2$ with phage containing central inserts previously considered to form two- and four-base loops (DAVISON and LEACH 1994b). The plaque-size measurements are consistent with the formation of loops containing two unpaired bases.

The experiments reported here describe the behavior of small numbers of trinucleotide repeats that affect the behaviour of a long palindrome *in vivo*. By contrast, the high levels of instability of trinucleotide sequences observed in human inherited disorders occur in long arrays of repeats. It is therefore appropriate to ask how our observations may relate to instability in long arrays. Two possible links exist.

1. It is known that both the kinetics of S-type cruciform extrusion *in vitro* (MURCHIE and LILLEY 1987; COUREY and WANG 1988; ZHENG and SINDEN 1988) and the inhibition of plaque formation by a long palindrome in an *sbcC* mutant (DAVISON and LEACH 1994a) are acutely center-dependent. This is understood to be because formation of small protocrucifom structures constitutes a kinetic barrier to cruciform extrusion and protocruciform formation is the rate-limiting step in the reaction. Similarly in a long tract of trinucleotide repeats, the formation of a small nucleating loop could constitute the rate-limiting step leading to the formation of a larger and more stable secondary structure. In our system such a secondary structure is formed by intrastrand base-pairing between the complementary "arms" of a long palindrome. In a long array of trinucleotides this may be a pseudo-hairpin or a more complex product.

2. Even a small secondary structure could result in a large change in number of repeats if the total tract of repeats is long. This possibility is exemplified by considering the cleavage and recombination model described in Figure 1. In this model the propensity of a repeat array to instability may not be determined by the size of the secondary structure but by the relationship of the array length to the minimal length of homology required to initiate homologous recombination. This minimal efficient processing segment (MEPS) (SHEN and HUANG 1986) is between 200 and 300 bp in mammalian cells (RUBNITZ and SUBRAMANI 1984; AYARES *et al.* 1986; LISKAY *et al.* 1987). For a repeat tract below this length, the pairing of the broken chromosome would have to rely on homology outside the repeat array. This would anchor the event and prevent significant changes in number of repeats. On the other hand, an array of repeats longer than the MEPS could recombine without external anchoring and lead to more frequent and variable changes in numbers. The observation that partial sequence divergence severely inhibits recombination in mammalian cells (WALDMAN and LISKAY 1987) would also account for the suppression of instability by imperfect repeats (CHUNG *et al.* 1993; HIRST *et al.* 1994; KUNST and WARREN 1994; SNOW *et al.* 1994).

We thank THORSTEN ALLERS for help in refining the methods and for many useful discussions, CHRIS JEFFREY for facilitating the image analysis and EWA OKELY for technical assistance. J. D. is supported by a studentship from Medical Research Council (MRC) Human Genome Mapping Project and this work is supported by MRC.

LITERATURE CITED

AYARES, D., L. CHEKURI, K.-Y. SONG and R. KUCHERLAPATI, 1986 Sequence homology requirements for intermolecular recombination in mammalian cells. Proc. Natl. Acad. Sci. USA **83:** 5199–5203.

CHALKER, A. F., D. R. F. LEACH and R. G. LLOYD, 1988 *Escherichia coli sbcC* mutants permit stable propagation of DNA replicons containing a long palindrome. Gene **71:** 201–205.

CHEN, X., S. V. S. MARIAPPAN, P. CATASTI, R. RATLIFF, R. K. MOYZIS *et al.*, 1995 Hairpins are formed by the single DNA strands of the fragile X triplet repeats: Structure and biological implications. Proc. Natl. Acad. Sci. USA **92:** 5199–5203.

CHUNG, M.-Y., L. P. W. RANUM, L. A. DUVICK, A. SERVADIO, H. Y. ZOGHBI *et al.*, 1993 Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. Nature Genet. **5:** 254–258.

COUREY, A. J., and J. C. WANG, 1988 Influence of DNA sequence and supercoiling on the process of cruciform formation. J. Mol. Biol. **202:** 35–43.

DAVISON, A., and D. R. F. LEACH, 1994a The effects of nucleotide sequence changes on DNA secondary structure formation in *Escherichia coli* are consistent with cruciform extrusion *in vivo*. Genetics **137:** 361–368.

DAVISON, A., and D. R. F. LEACH, 1994b Two-base DNA hairpin-loop structures *in vivo*. Nucleic Acids Res. **22:** 4361–4363.

FRY, M., and L. A. LOEB, 1994 The fragile X syndrome $d(CGG)_n$ nucleotide repeats form a stable tetrahelical structure. Proc. Natl. Acad. Sci. USA **91:** 4950–4954.

GACY, A. M., G. GOELLNER, N. JURANIC, S. MACURA and C. T. MCMURRAY, 1995 Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell **81:** 533–540.

GIBSON, F. P., D. R. F. LEACH and R. G. LLOYD, 1992 Identification of *sbcD* mutations as cosuppressors of *recBC* that allow propagation of DNA palindromes in *Escherichia coli* K-12. J. Bacteriol. **174:** 1222–1228.

HASTINGS, P. J., 1988 Recombination in the eukaryotic nucleus. BioEssays **9:** 61–64.

HILBERS, C. W., H. A. HEUS, M. J. P. VAN DONGEN and S. S. WIJMENGA, 1994 The hairpin elements of nucleic acid structure: DNA and RNA folding. Nucleic Acids Mol. Biol. **8:** 56–104.

HIRST, M. C., P. K. GREWAL and K. E. DAVIES, 1994 Precursor arrays for triplet repeat expansion at the fragile X locus. Hum. Mol. Genet. **3:** 1553–1560.

IMBERT, G., C. KRETZ, K. JOHNSON and J.-L. MANDEL, 1993 Origin of the expansion mutation in myotonic dystrophy. Nature Genet. **4:** 72–76.

JANSEN, G., P. WILLEMS, M. COERWINKEL, W. NILLESEN, H. SMEETS *et al.*, 1994 Gonosomal mosaicism in myotonic dystrophy patients: involvement of mitotic events in $(CTG)_n$ repeat variations and selection against extreme expansion in sperm. Am. J. Hum. Genet. **54:** 575–585.

JEFFREYS, A. J., K. TAMAKI, A. MACLEOD, D. G. MONCKTON, D. L. NEIL *et al.*, 1994 Complex gene conversion events in germline mutation at human minisatellites. Nature Genet. **6:** 136–145.

KUNST, C. B., and S. T. WARREN, 1994 Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. Cell **77:** 853–851.

LEACH, D. R. F., 1994 Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. BioEssays **16:** 893–900.

LEVINSON, G., and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

LISKAY, R. M., A. LETSON and J. STACHELEK, 1987 Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. Genetics **115:** 161–167.

MITAS, M., A. YU, J. DILL, T. J. KAMP, E. J. CHAMBERS *et al.*, 1995 Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: $(CTG)_{15}$. Nucleic Acids Res. **23:** 1050–1059.

MURCHIE, D. I. H., and D. M. J. LILLEY, 1987 The mechanism of cruciform formation in supercoiled DNA: initial opening of central basepairs in salt-dependent extrusion. Nucleic Acids Res. **15:** 9641–9654.

NASMYTH, K. A., 1982 Molecular genetics of yeast mating type. Annu. Rev. Genet. **16:** 439–500.

RESNICK, M., 1976 The repair of double strand breaks in DNA: a model involving recombination. J. Theor. Biol. **59:** 97–106.

RICHARDS, R. I., and G. R. SUTHERLAND, 1994 Simple repeat DNA is not replicated simply. Nature Genet. **6:** 114–116.

RITCHIE, R. J., S. J. L. KNIGHT, M. C. HIRST, P. K. GREWAL, M. BOBROW *et al.*, 1994 The cloning of *FRAXF*: trinucleotide repeat expansion and methylation at a third fragile site in distal Xqter. Hum. Mol. Genet. **3:** 2115–2121.

RUBNITZ, J., and S. SUBRAMANI, 1984 The minimum amount of homology required for homologous recombination in mammalian cells. Mol. Cell. Biol. **4:** 2253–2258.

SHEN, P., and H. V. HUANG, 1986 Homologous recombination in *Escherichia coli:* dependence on length and homology. Genetics **112:** 441–457.

SINDEN, R. R., and R. D. WELLS, 1992 DNA structure, mutations and human genetic disease. Curr. Opin. Biotechnol. **3:** 612–622.

SMITH, S. S., A. LAAYOUN, R. G. LINGEMAN, D. J. BAKER and J. RILEY, 1994 Hypermethylation of telomere-like foldbacks at codon 12 of the human c-Ha-*ras* gene and the trinucleotide repeat of the *FMR-I* gene of fragile X. J. Mol. Biol. **243:** 143–151.

SNOW, K., D. J. TESTER, K. E. KRUKEBERG, D. J. SCHAID and S. N. THIBODEAU, 1994 Sequence analysis of the fragile X trinucleotide repeat; implications for the origin of the fragile X mutation. Hum. Mol. Genet. **3:** 1543–1551.

STREISINGER, G., Y. OKADA, J. EMRICH, J. NEWTON, A. TSUGITA *et al.*, 1966 Frameshift mutations and the genetic code. Cold Spring Harbor Symp. Quant. Biol. **31:** 77–84.

THALER, D. S., M. M. STAHL and F. W. STAHL, 1987 Tests of the double-strand break repair model for red-mediated recombination of phage λ and plasmid λdv. Genetics **116:** 501–511.

WALDMAN, A. S., and R. M. LISKAY, 1987 Differential effects of base-pair mismatch on intrachromosomal versus extrachromosomal recombination in mammalian cells. Proc. Natl. Acad. Sci. USA **84:** 5340–5344.

WILLEMS, P. J., 1994 Dynamic mutations hit double figures. Nature Genet. **8:** 213–215.

ZHENG, G., and R. R. SINDEN, 1988 Effect of base composition in the center of inverted repeated DNA sequences on cruciform transitions in DNA. J. Biol. Chem. **263:** 5356–5361.

Communicating editor: R. MAURER

J. Mol. Biol. (1998) 275, 3–16

JMB

REVIEW

Secondary Structures in d(CGG) and d(CCG) Repeat Tracts

John M. Darlow and David R. F. Leach*

Institute of Cell and Molecular Biology, University of Edinburgh, King's Buildings Edinburgh EH9 3JR, UK

**Corresponding author*

Several studies have been made to elucidate the nature of secondary structures in the single strands of d(CGG)·d(CCG) repeat tracts but with conflicting conclusions. Here, we review this work and attempt to come towards consensus. Some investigators find that the G-rich strand forms hairpins. Of these, some conclude that pairing is in the alignment d(GGC)·d(GGC) with two Watson-Crick bonds and one G·G bond per duplex repeat, others conclude that the alignment is d(GCG)·d(GCG) with two G·G bonds and one C·C bond per duplex repeat. Others find quadruplex formation and conclude that this is in the latter alignment with two G_4-quartets per quadruplex repeat and C·C bonds. We investigate why these different results were obtained and conclude that quadruplexes are likely to form under physiological conditions. We argue that they are probably bonded in the alignment d(GGC)·d(GGC) with one G_4-quartet and two C·G·C·G· quartets per quadruplex repeat. The C-rich strand does not appear to form quadruplexes under physiological conditions but forms hairpins. Apparently, short hairpins adopt the alignment d(CCG)·d(CCG) with mismatched cytosine residues stacked into the helix but with 15 or more repeat units, the dominant form is a distorted hairpin aligned as d(GCC)·d(GCC) with unpaired cytosine residues possibly turned outwards and stacked in the minor groove.

Keywords: d(CGG)·d(CCG) repeat tracts; DNA hairpins; quadruplexes; homoduplex alignment; annealing conditions

Introduction

Expansions of d(CGG)·d(CCG) repeat tracts have been found to be the basis of all the folate-sensitive human fragile chromosome sites so far sequenced (Sutherland & Richards, 1995) and were first brought to light by molecular genetic investigation of fragile-X syndrome. The formation of unusual secondary structures by either or both of the DNA strands is thought to be involved in the genetic instability of this sequence. In order to alleviate the apparent confusion over the number of possible hairpins that might be formed by single strands of d(CGG)·d(CCG) repeats we illustrate them all in Figure 1 and will refer constantly to this scheme. We define each possible alignment in terms of the frame in which the sequence 5′-3′ is the same on both sides. In alignments with frames 1 and 2, two out of three bases in each trinucleotide are involved in Watson-Crick base-pairs and the remaining base is in a C·C or G·G mispair depending upon the strand. Frame 1 (d(CGG)·d(CGG) and d(CCG)·d(CCG); alignment (a) of Mitas *et al.* (1995) and Yu *et al.* (1997) but alignment B of Gao *et al.* (1995)) is akin to the pairing $d(CAG)_n \cdot d(CAG)_n$ and $d(CTG)_n \cdot d(CTG)_n$ that may occur in the single strands of the repeat sequence found in the genes responsible for several human inherited disorders. Frame 2 (d(GGC)·d(GGC) and d(GCC)·d(GCC); alignment (b) of Mitas *et al.* (1995) and Yu *et al.* (1997) but alignment A of Gao *et al.* (1995)) is akin to the pairing $d(GAC)_n \cdot d(GAC)_n$ and $d(GTC)_n \cdot d(GTC)_n$ that could occur in the two strands of the other possible trinucleotide repeat, which consists of the same bases as $d(CAG)_n \cdot d(CTG)_n$ but that has not been found in lengths of more than five repeat units in the human genome (Gacy *et al.*, 1995).

In Frame 3 (d(GCG)·d(GCG) and d(CGC)·d(CGC); alignment C of Gao *et al.*, 1995) there is no Watson-Crick base-pair but attention has been drawn to this alignment because of the possibility that at least the G-rich strand might be able to form a quadruplex structure held together by G_4-

0022-2836/98/010003-14 $25.00/0/mb971453

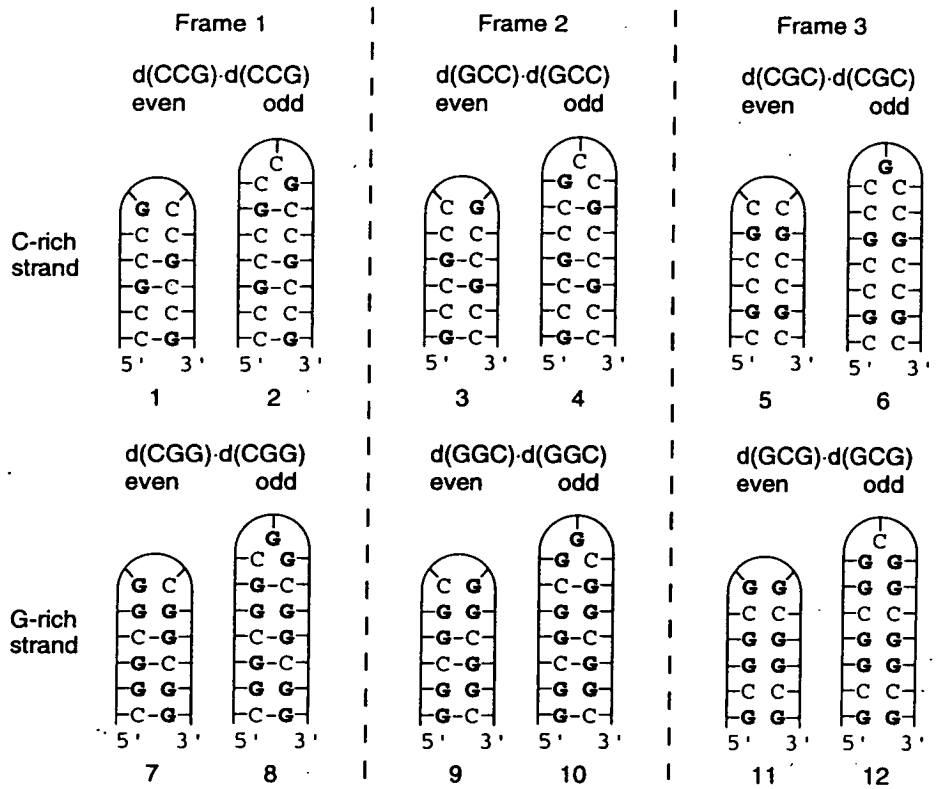


Figure 1. All the possible types of hairpin loops that might be formed by single strands of d(CGG)·d(CCG) repeats. Only Watson-Crick bonds are shown but other workers have found evidence of G·G bonds and C·C bonds *in vitro* (see the text). N.B. The above classification is based upon the alignment of the two sides of the hairpin and the presence of an odd or even number of unpaired bases in the loop. It does not depend upon the actual number of unpaired bases in the loop, length of the stem or the 5′-base of the sequence. We define the alignment by the frame in which the sequence 5′-3′ is the same on both sides of the stem. For all three alignments some workers have postulated that long hairpins might fold over to form unistrand quadruplexes.

"quartets" (see Figure 4(c)). There has been much interest in the potential for formation of such structures by G-rich sequences that occur in telomeres (Venczel & Sen, 1993; Williamson, 1994; Kettani *et al.*, 1995). It has been shown that such structures can form *in vitro* from four DNA strands, from two hairpins, and from a single strand. It has not been proven that they occur *in vivo* with telomeres but there is evidence of quadruplex formation *in vivo* with a G-rich sequence upstream of the human insulin gene (Hammond-Kosack *et al.*, 1992). These quadruplexes require the presence of cations for their formation. In the case of monovalent ions, a single cation sits between two G_4-quartets in an octahedral complex with the carbonyl groups of the guanine residues. The divalent cations have been thought to promote DNA structures with tight helices by acting as counterions between the phosphate oxygen atoms of adjacent backbones (Venczel & Sen (1993) and references therein) but Venczel & Sen (1993), noting the similarities of the stabilizing orders of monovalent and divalent ions, raised the possibility that the divalent ions might also be complexed between the guanine quartets. The divalent cations achieve their effect with about two orders of magnitude lower concentrations than the monovalent cations (Venczel & Sen, 1993; Lee, 1990) but the best effect achieved by any ion is dependent upon the van der Waals radius of the hydrated ion, not just its charge and concentration, and K^+ fits best into the octahedral cage. The divalent magnesium ion is only about the same size as that of the monovalent lithium ion. It has a larger effect upon stability because of its higher charge but still has a lesser effect than Na^+. The divalent calcium ion is similar in size to that of Na^+. Thus the order of stabilizing ability for the main intracellular cations is $K^+ > Ca^{2+} > Na^+ > Mg^{2+}$ (Hardin *et al.*, 1992).

For d(CGG) repeats it has been suggested that hairpins might form in frame 3, held together by G·G Hoogsteen bonds and possibly by C·C bonds in addition. It has been suggested that long hairpins of this type might fold over onto themselves to form quadruplexes. A diagram of such a structure is shown in Figure 2(a). Mitas *et al.* (1995) have suggested that hairpins of the G-rich strand in frame 1 or 2 might similarly fold to form quadruplexes, and that these might contain C·G·C·G· quartets. In either frame there would be two C·G·C·G· quartets to every one G_4-quartet. Diagrams of examples of these quadruplexes are given in Figure 2(b) and (c). It has now been shown that a quadruplex containing C·G·C·G· and G_4-quar-

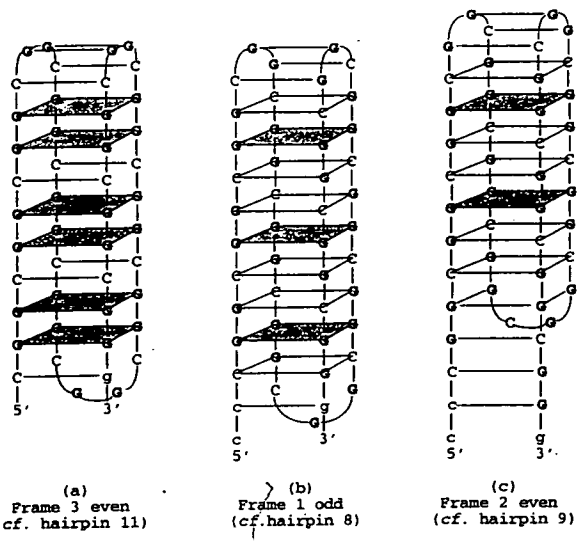


Figure 2. Examples of quadruplexes in the three alignments of pairing. (b) and (c) The two quadruplexes suggested by Mitas *et al.* (1995) for $d(CGG)_{15}$ redrawn to give an impression of three dimensions. The bases indicated by small letters are flanking bases of their construct. (a) A quadruplex in frame 3 that might be formed by the same sequence, $d(CGG)_{15}$ (and one of the flanking bases). The front left strands of all three structures are aligned. G_4-quartets are shaded to emphasize the different patterns and C·G·C·G· quartets are indicated by unshaded rhomboids.

tets can form *in vitro* (Kettani *et al.*, 1995); in that study the cation used was Na^+.

By a careful examination of all the data now available on secondary structure in the single strands of d(CGG)·d(CCG) repeats we have sought to discern the natures of the most stable structures formed by each strand. Because different workers have come to opposing conclusions upon various points, we have necessarily had to disagree with some and to reinterpret their data. We have divided our discussion roughly into sections on each of the two strands but cannot do this completely as inevitably some work compares the two.

The G-rich Strand: Frame 1, 2 or 3?

Investigations giving evidence for alignment in frame 1 or 2

Mitas *et al.* (1995) considered two alignments, frame 1 (alignment (a)) and frame 2 (alignment (b)). They used an oligonucleotide (excised from a plasmid) containing the sequence $dCC(CGG)_{15}G$. They considered the possibilities of hairpins 8 and 9, both with and without G·G bonds, and quadruplexes formed by either of the two hairpins folded over on itself and bonded by one G quartet and two C·G·C·G· quartets per repeat (Figure 2(b) and (c)). By methylation of the self-annealed oligonucleotide with dimethyl sulphate (DMS) followed by cleavage at the modified bases and electrophoresis under denaturing conditions it was possible to investigate the nature of the secondary structure. DMS methylates the N7 positions of guanine residues. In G·C bonds, and in the C·G·C·G· quartet arrangement described by Mitas *et al.* (1995) the N7 positions are not involved in hydrogen bonding; in G·G bonds the N7 position of one of the guanine residues is involved, and in G_4-quartets the N7 positions of all four residues are hydrogen-bonded and thereby protected from methylation. No guanine residue was completely protected under a wide range of conditions, which was deemed to rule out quadruplex structure. Residues in the stem of a hairpin or quadruplex are protected relative to those in unpaired loops. A hairpin has one loop of unpaired bases, whereas a unimolecular quadruplex has three (see Figures 2 and 5). DMS and P_1 nuclease, which also attacks these loops, both indicated that there was only one loop, again suggesting hairpin structure. Relative reactivities of bases in the loop decreased with increasing concentration of KCl, indicating that potassium stabilizes the structure. A melting study showed that the t_m of $d(CGG)_{15}$ is 27 deg. C higher than that of $d(CTG)_{15}$, from which the authors concluded that G·G base-pairs contribute a significant amount of stability to the hairpin. Relative methylation of the two G residues of the GpG dinucleotide can distinguish between the alignments because in frame 1 it is the 5′ G residues that are involved in G·G bonds, whereas in frame 2 it is the 3′ G residues. This study indicated that at KCl concentrations of $\geqslant$200 mM the hairpins were all aligned in frame 2 (hairpin 9) but that at concentrations of $\leqslant$100 mM this hairpin must be in equilibrium with another structure that offers less methylation protection to the 3′ G residues. This could either be the same hairpin but without G·G bonds or it could be a hairpin 8 structure with G·G bonds.

An NMR study (Chen *et al.*, 1995; Mariappan *et al.*, 1996) concluded that pairing of short $d(GGC)_n$ oligonucleotides was in frame 2 under all conditions tested, and that the mismatched G residues were strongly paired and stacked with the neighbouring G·C pairs. The authors were unable to examine loop structure by NMR because of the predominance of the homoduplex $d(GGC)_n$ over hairpin at the DNA concentrations they used. They investigated hairpin folding by electrophoresis in non-denaturing gels. They plotted percentage of hairpin against duplex DNA for $n = 5$, 6, 7 and 11, and found that at all salt concentrations tested $d(GGC)_5$ showed a greater tendency to hairpin formation than did $d(GGC)_6$, and they attributed this to the number of bases in the loop. The result indicates that if pairing in the stem of the hairpins is the same as that of the duplex, i.e. frame 2, then a hairpin 10 structure is preferred over hairpin 9. Another NMR study of these repeats (Zheng *et al.*, 1996) also found in favour of alignment in frame 2. Its deduction is made from observations of $d(CGG)_3$ with the second, third or sixth residue

substituted by inosine, with the G residues of the other triplets unsubstituted. With the I2 substitution, an imino proton resonance characteristic of an I·C bond was found but with either of the I3 and I6 substitutions the resonance was that of a non-hydrogen-bonded inosine residue.

The NMR studies agree about the alignment of self-annealing of the G-rich strand but disagree about the mispaired G residues. Zheng *et al.* (1996) found very broad imino-proton resonances for these bases and could not detect amino proton resonances or intra-residue NOEs. They concluded that the residues were unpaired and very mobile, most likely undergoing dynamic exchange among various glycosidic conformational isomers. Mariappan *et al.* (1996) found an imino-proton peak corresponding to G·G bonds with broad resonances either side, which they attributed to the minor hairpin population (presumably the unpaired G residues in the loop). They also found a strong NOE connecting the G·C and G·G imino protons and concluded that the mispaired G residues were strongly base-paired through the imino protons and stacked with the neighbouring G·C pairs.

Investigations interpreted as showing alignment in frame 3

Sinden & Wells (1992) suggested that a single strand of $d(CGG)_n$ might form a hairpin aligned in frame 3 with G·G bonds, quoting work referring to quadruplex DNA with guanine quartets. Fry & Loeb (1994) examined the possibility of quadruplex formation with short oligonucleotides. They melted them and then incubated them for up to 90 hours at 4°C and found that at pH 8 in 200 mM KCl $d(CGG)_4$ and $d(CGG)_5$ would form species that were slow-moving on non-denaturing gels if the cytosine residues were methylated, and that $d(CGG)_7$ would do so even if not methylated, but such species were not formed by the corresponding C-rich oligonucleotides. They then investigated the dependence of the formation of the slow-moving complexes on the presence of various cations, examined their kinetics and stoichiometry of formation and resistance to methylation by DMS and concluded that the G-rich oligonucleotides formed quadrimolecular quadruplexes. At this time the possibility of quadruplex formation by $d(CGG)_n$ in frames 1 or 2 had not been suggested. The authors assumed that bonding was in frame 3 and none of their experiments could have distinguished the frame. They also apparently assumed that the strands would be parallel (as opposed to antiparallel). Like Lee (1990), who studied other quadruplexes, they found that there was an optimum concentration of Mg^{2+}, 4 mM, for quadruplex formation by $d(^{5m}CGG)_5$. The maximum percentage of the total DNA in the slow-moving complex, 12.6%, was less than that achieved with any of the other ions they tried. They plotted percentage of quadruplex formed after a fixed time for different concentrations of K^+, Na^+ and Li^+, and percentage formed at fixed ion concentrations after different time periods. In each case K^+ fostered a much higher percentage than Na^+. Sen & Gilbert (1990) found that with G-rich oligonucleotides that were capable of forming quadruplexes both from two hairpins and from four strands, K^+ induced the rapid formation of bimolecular quadruplexes that were so stable that they would not unfold to allow quadrimolecular ones to form. To achieve four-stranded structures it was necessary to use Na^+, which did not stabilize either structure as well, and then K^+ could be substituted after the four-stranded quadruplexes had formed. Thus the finding of better quadruplex formation with K^+ by Fry & Loeb (1994) suggests that either they were measuring bimolecular quadruplex or that the oligonucleotides that they were using had little tendency to form hairpins under their conditions. The latter explanation would agree with the findings of Chen *et al.* (1995) and Mariappan *et al.* (1996). The surprise was that the greatest percentage of the slow-moving complex was reached with Li^+ (about 55% at 400 mM Li^+ in 49 hours). Fry & Loeb (1994) suggested that this might indicate that the guanine tetrads (in these $d(CGG)_n$ quadruplexes) might be packed more tightly than in quadruplexes formed from short guanine tracts dispersed among non-guanine sequences.

Fry and colleagues (Nadel *et al.*, 1995) went on to examine the possibility of unimolecular quadruplex formation by studying fast-moving electrophoretic species. This time, in addition to very short oligonucleotides, they included $d(GCG)_8$, $d(GCG)_{11}$ and other forms of some of these in which one, two or three of the bases were replaced by thymine in each of the places where the unpaired loops would be if the molecule folded into hairpins or quadruplexes. It is clear that they assumed that the alignment would be in frame 3, because the thymine residues were placed in positions such that only in this alignment would the pairing be unaffected by their presence (whether the structure was a quadruplex or a hairpin). From electrophoretic, kinetic and UV-cross-linking studies they concluded that unimolecular secondary structures were formed. However, three pieces of evidence suggested that the structures were hairpins and not quadruplexes. Firstly, their formation was not dependent upon cations. Secondly, they concluded that all the guanine residues were modified by DMS. Thirdly, modification with diethyl pyrocarbonate (DEPC, which reveals unpaired purine residues) showed that the $d(GCG)_n$ sequences contained only one loop of unpaired bases.

The authors then deduced the structures of the hairpins from DEPC and $KMnO_4$ modification results. The results for a substituted form of $d(GCG)_{11}$, $d(GCG)_2T_3(GCG)_2T_3(GCG)_2T_3(GCG)_2$ show minimal cleavage at any of the G residues. If there was folding in frame 1 or 2 there would be some unpaired guanine residues opposite thymine residues. Thus it appears that this molecule folds in

frame 3 with all of the guanine residues in G·G bonds and three loops of unpaired thymine residues. It is seen, however, that this is only because the positioning of the thymine residues is such that it would be energetically disadvantageous to pair in any other way because the unsubstituted $d(GCG)_{11}$ does not pair in this frame. The authors concluded that $d(GCG)_{11}$ paired in frame 1 with a one-base 3′ overhang (a hairpin 7 structure) or possibly in frame 2 with a two-base 3′ overhang (hairpin 10). This interpretation is marred by the fact that they have interpreted bands on their autoradiograms as cleavage at the wrong ends of the molecules. Reading from the correct end of the molecule, the unpaired loop appears to be 5′GCGG3′. This corresponds very nicely with the loop of hairpin 9 and indicates frame 2. The loop is an even-membered one and there would be a one-base 5′ overhang. The data presented by Mariappan *et al.* (1996) suggest that an odd-membered loop may be more stable in frame 2 but with this oligonucleotide this structure would have required a four-base 5′-overhang or a two-base 3′ overhang, both of which would probably be energetically less favourable than the one the autoradiogram appears to show.

The reason why Nadel *et al.* (1995) failed to find unimolecular quadruplexes is given by the work of Usdin & Woodford (1995). The latter cloned $d(CGG)_n$ and $d(GCC)_n$ tracts in M13mp18 to make single-stranded templates, then polymerized complementary strands (at pH 9.3) from a primer outside the repeat tract and examined the results by electrophoresis. They found that, with G-rich templates only, there were strong blocks to synthesis of the new strand with all of five polymerases tried. These occurred only if $n \geqslant 13$. They were at the 3′ end of the template so could not be due to formation of triplex DNA between the template and nascent strand. Arrest was independent of template concentration, suggesting an intramolecular structure. For the activity of the polymerase, Mg^{2+} was of course present, as 2.5 mM $MgCl_2$, but the blocks were K^+-dependent. Little if any DNA synthesis arrest was seen in the absence of a monovalent cation or when NaCl, NH_4Cl, RbCl or CsCl were used in place of KCl. With KCl it still occurred even after prolonged incubation of the DNA at 85°C before addition of a heat-stable polymerase. Usdin & Woodford (1995) found that arrest of synthesis was eliminated by replacement of the second guanine residue of each of the last four CGG triplets in a template containing $d(CGG)_{16}C$ with 7-deazaguanine in which N7 is not free to take part in hydrogen bonding, thus ruling out a hairpin containing only G·C bonds. Since the substitution of only four of the 32 guanine residues was required to abolish arrest, they considered it unlikely that the structure was a hairpin, either with only G·G base-pairs (i.e. frame 3) or a mixture of C·G and G·G bonds (i.e. frame 1 or 2), because the N7 of only 50% or 33%, respectively, of guanine residues would have to be involved in these cases.

Electrophoresis and chemical probing was then performed on a 90-mer oligonucleotide containing $d(CGG)_{20}$ and in these experiments no Mg^{2+} was present. In a denaturing gel the molecule ran, as expected, well behind a 69-mer marker. In a non-denaturing polyacrylamide gel containing Tris-borate/EDTA buffer and no added salt it ran at the same speed as the marker, i.e. much faster than before, and under these conditions it would be expected to be a hairpin. In 40 mM LiCl it ran in the same place but in 40 mM KCl it ran faster. This fast mobility was eliminated by methylation of the guanine residues by DMS. From all this evidence the authors concluded that the structure was some sort of intrastrand quadruplex. They found that in the presence of K^+ the oligonucleotide (at the same concentration) was almost completely protected from methylation of the N7 positions by DMS. This is a very impressive result, showing much greater methylation protection than was found by Nadel *et al.* (1995) with $d(CGG)_{11}$ or by Mitas *et al.* (1995) with $d(CGG)_{15}$. Usdin & Woodford (1995) modified their oligomer with bromoacetaldehyde (BAA) followed by formic acid or DMS and then cleaved with pyrrolidine to detect unpaired cytosine residues and found only the 11th cytosine of $d(CGG)_{20}$ to be unpaired whether potassium was present or not. The authors pointed out that in the absence of potassium this would be consistent with a hairpin 12 structure (Figure 1). It could actually fit with hairpins 7, 8, 9 or 10 too. In the presence of K^+, they concluded that the single unpaired C and almost complete protection of all the G residues must indicate a quadruplex in frame 3. This is illustrated by Figure 5(a).

Triad DNA

For completeness, we should mention that Kuryavyi & Jovin (1995a,b) proposed a structure for d(CAG) and d(CGG) repeats that they called triad-DNA. In this, the trinucleotide homoduplex forms an antiparallel double helix but alternately two adjacent bases on one strand are paired with one base on the other strand and next to that two bases on the second strand are paired with one on the first strand, the sugar-phosphate backbones having to make unusual turns to achieve this. Molecular mechanics calculations predicted that this structure should be more stable than a duplex with G·G mismatches. No evidence for the formation of triad-DNA with these repeats has been found, though two single (T·A)·A triads have been demonstrated sandwiching two G_4-quartets in a quadruplex of d(TTAGG) pentanucleotides (Kettani *et al.*, 1997).

Studies relevant to the frame of pairing of quadruplexes

The postulation and computer modelling of quadruplexes containing C·G·C·G· quartets by Mitas *et al.* (1995) has already been mentioned.

Two other laboratories discovered that they could actually exist. Leonard *et al.* (1995) crystallized the heptanucleotide d(GCATGCT) and were surprised to find that, instead of a Watson-Crick duplex, it formed a bimolecular quadruplex composed of two hairpins. These were held together by two G·C·G·C· quartets. In addition, of the bases in the loops, the thymine residues were turned outwards but the adenine residues formed an A·A bond diagonally across the structure. It is notable that this quadruplex has no G_4-quartets. The crystals were formed by vapour diffusion from sitting drops of a starting suspension of about 5 mg/ml of the heptanucleotide in 50 mM lithium cacodylate, 50 mM $MgCl_2$.

Kettani *et al.* (1995) set out to investigate the possibility of quadruplex formation by the G-rich strand of d(CGG)·d(CCG) repeats with the oligonucleotide $dGCGGT_3GCGG$. It might have formed hairpins and quadruplexes in any alignment (Figure 3) but with such a short sequence it is not surprising that the authors found that it formed hairpins with no overhangs, which is actually frame 2. By NMR they found that these hairpins associated to form quadruplexes, using all the G and C bases, with two G·C·G·C· quartets sandwiched between two G_4-quartets. While this study showed that a frame 2 quadruplex is possible, the sequence was so short that the energy cost of an overhang might have determined the frame. To date, no NMR study has been published on longer $d(CGG)_n$ quadruplexes.

The C·G·C·G· quartets of these groups are not the same. There are two main types of C·G·C·G· quartet. Both are formed by the association of two C·G base-pairs in opposite directions but in one, which we will call type 1 (Figure 4(a)) the major groove sides are facing one another and in the other, type 2 (Figure 4(b)) the minor groove sides are facing one another. For the type 1 quartet, different variants have been proposed. In one of these, the two base-pairs are bonded to each other *via* N4 of the cytosine residues and O6 of the guanine residues (bonds labelled 1 in Figure 4(a)). This was the scheme derived by Mitas *et al.* (1995) from computer modelling and had been proposed in the 1960s (Löwdin, 1964; Kubitschek & Henderson, 1966) for association of two Watson-Crick double helices in models of DNA replication. In this arrangement the N7 positions of the guanine residues are not involved. Another variation was discovered by O'Brien (1967) by X-ray crystallography of graphite-like crystals of a 1:1 complex of 9-ethylguanine and 1-methylcytosine (no sugar or phosphate group involved). In this, N4 of the cytosine residues were hydrogen-bonded to N7 of the guanine residues of the opposite C·G pair (bonds labelled 2 in Figure 4(a)). Ketani *et al.* (1995) claim that there is a bifurcated hydrogen bond (i.e. 1 and 2) in the C·G·C·G· quartets in their quadruplex. However, McGavin (1971) drew this arrangement, half-way between the first two bonding schemes, and commented that the NH—N distance was probably too long for the O'Brien scheme, and that the angle between NH and NO was probably too large for the alternative scheme and from his discussion he seemed to consider that the bonding had to be one or the other. Subsequently, Williams *et al.* (1989) have confirmed the finding reported by O'Brien (1967). Williams *et al.* (1989) used 2'-deoxynucleosides of guanine and cytosine, substituted at both ribosyl hydroxyl groups with triisopropylsilyl groups and disolved them in chloroform-*d*, a low dielectric solvent that promotes the formation of hydrogen bonds as opposed to stacking. They then examined the result by NMR and, like O'Brien, found C·G·C·G· quartets with bonds 2 only. The quadruplex described by Leonard *et al.* (1995) has type 2 quartets in which the two C·G pairs are bonded to each other *via* O2 of the cytosine residues and N2 of the guanine residues. Both types of quadruplex have two wide grooves and two narrow grooves but in the second type the difference between the groove sizes is much larger. Also, the type 1 quartet is roughly planar, whereas in the type 2 quartet the two C·G pairs are tilted at ~30° to one another about an axis going through the C·G bonds. G_4-quartets (Figure 4(c)) are roughly planar and roughly square (though quadruplexes with G_4-quartets do have narrow and wide grooves), so if CGG repeats do form quadruplexes in frame 2 the C·G·C·G· quartets would have to be type 1 in order to stack on the G_4-quartets.

In contrast to the above studies, Chen (1995) found pairing in frame 3. He used absorbance, circular dichroic and gel measurements to monitor the aggregation of $d(CGG)_4$ oligonucleotides beyond a starting mixture of hairpins and linear duplexes. He found that kinetics were extremely

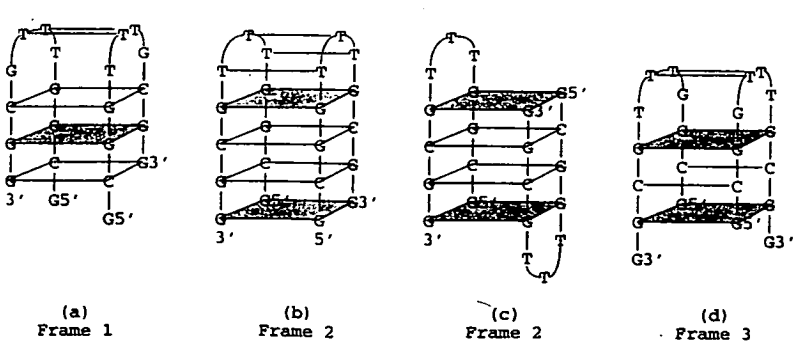


Figure 3. Some of the structures that the oligonucleotide studied by Kettani *et al.* (1995) might have formed. The actual structure was (c) (see the text).

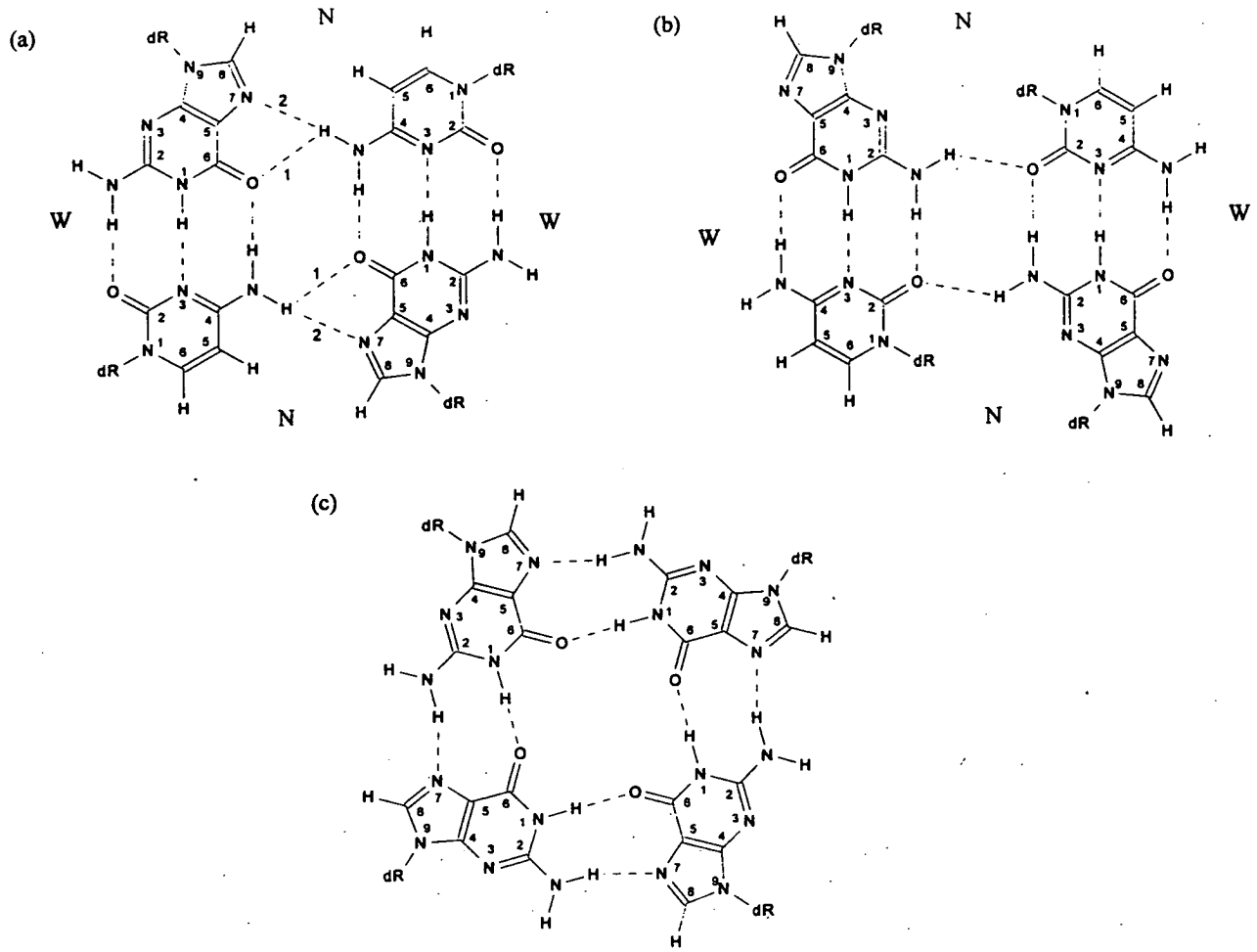


Figure 4. (a) Type 1 C·G·C·G· quartet; (b) type 2 C·G·C·G· quartet; (c) G_4-quartet. W, wide grooves; N, narrow grooves.

slow at pH 8, taking about ten days to reach equilibrium with 2 M KCl at room temperature. A pH of 5·4 and >0·8 M KCl were required to observe the onset of aggregation at 20°C within one day and the formation was of parallel-stranded quadruplexes. Chen (1995) further concluded that after these formed, the G_4-quartets on either side of the cytosine residues stacked together, leaving the cytosine residues protruding outwards from the condensed tetramer and paired with the cytosine residues of neighbouring quadruplexes to form larger multiplexes.

Attempts at resolution of the confusion

Several questions arise at this stage. What is the range of conditions under which quadruplexes will form with $d(GGC)_n$ tracts and why did some groups find evidence of quadruplexes and others not? Does the sequence form both frame 3 and frame 2 quadruplexes under different conditions? Does it form quadruplexes under physiological conditions and, if so, what kind are they?

First, why did Mitas *et al.* (1995) not detect any quadruplex formation with 15 triplets yet Usdin & Woodford (1995) appeared to find it with as few as 13 triplets? The conclusions reached by Mitas *et al.* (1995) hinged upon the methylation conditions but comparison of the methods shows that the answer does not lie there. It appears to lie in the annealing conditions. After melting their DNA and immediately before adding DMS, Usdin & Woodford (1995) incubated at 37 or 55°C for five minutes, during which the quadruplexes evidently formed, while Mitas *et al.* (1995) put theirs on ice for five minutes and only hairpins resulted.

It seems to us that Mitas *et al.* (1995) may indeed have had a monomolecular quadruplex when they estimated the melting point of $d(CGG)_{15}$ secondary structure to be about 75°C. On non-denaturing gels, single-stranded $d(CGG)_{15}$ ran ahead of single-stranded $d(CTG)_{15}$, suggesting to us that it formed a more compact structure, possibly a quadruplex. We note that the DNA was pre-annealed for five minutes at 25°C rather than on ice. The gels contained Tris-borate/EDTA buffer (pH 8·5) with no added salt but it appears that the buffer in which the DNA was suspended may have contained enough Na^+ for quadruplex formation. Three observations are consistent with our interpretation.

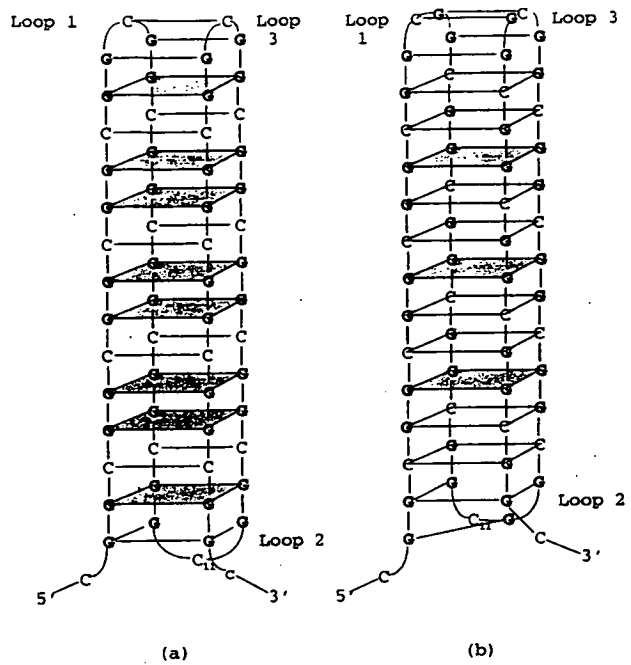


Figure 5. Two alternative possible structures for quadruplexes formed by the sequence $d(CGG)_{20}C$. (a) One of the frame 3 structures envisaged by Usdin & Woodford (1995); (b) one of the possible structures aligned in frame 2.

Firstly, though we could not find any reference to a t_m measurement of a unimolecular antiparallel quadruplex of the right size, we note that Sen & Gilbert (1990) found that for tetramolecular parallel-stranded quadruplexes of oligonucleotides of the sequence dTGGGGAGCTGGGGT the melting point in the presence of K^+ was over 95°C but in the presence of Na^+ only was between 75 and 80°C. This quadruplex had nine G_4-quartets and so might have about the same stability as a quadruplex of $d(CGG)_{15}$ (Figure 2). Furthermore, bimolecular antiparallel quadruplexes with eight G_4-quartets have been reported to have a t_m of about 55°C in Na^+ solution (Hardin *et al.*, 1991), which is all the more reason to think that the structure reported by Mitas *et al.* (1995) was not a hairpin. Secondly, Mitas *et al.* (1995) found that when the cytosine residues of their $d(CGG)_{15}$ oligonucleotide were C5-methylated the melting point was about 83°C. C5-methylation of cytosine residues increases the stability of quadruplexes with C·C bonds and G_4-quartets, probably by improving base stacking (Hardin *et al.*, 1993); it increased the stability of the quadruplexes reported by Fry & Loeb (1994) and appears likely to stabilize quadruplexes with a mixture of C·G·C·G· quartets and G_4-quartets also by improving stacking (Kettani *et al.*, 1995). Thirdly, Smith *et al.* (1994) produced a supporting result. They performed electrophoresis of $d(CCG)_{15}$ and $d(GGC)_{15}$ and their counterparts with inosine substituted for guanine, $d(CCI)_{15}$ and $d(IIC)_{15}$. (Telomeric sequences with inosine substituted for guanine do not cohere (Henderson *et al.*, 1990; Acevedo *et al.* 1991).) In a non-denaturing gel, $d(CGG)_{15}$ ran more than twice as far ahead of the I-substituted molecules as did $d(CCG)_{15}$. We considered that this might indicate unimolecular quadruplex formation and Steven Smith (personal communication) has kindly supplied us with their exact experimental details. After melting, the DNA was annealed in 100 mM NaCl (pH 7·4) at 60°C for ten minutes and at room temperature (22°C) for ten minutes. Electrophoresis was performed in Tris-borate/EDTA buffer (pH 8.3) with no added salt but from their experience they have concluded that when these structures are formed during annealing they are stable during electrophoresis. The absence of smears of label down the gel reassured them about this. They agree with us in our suspicion of quadruplex formation. If we are right, we have to explain why Usdin & Woodford (1995) did not detect blocks to DNA synthesis on a $d(CGG)_{20}$ template in the presence of Na^+. This is presumably because their assay was performed using *Taq* polymerase at 72°C, which may be too close to the melting temperature of the quadruplex structure in the presence of Na^+.

We now come to the NMR investigations of $d(GGC)_n$ oligonucleotides. Both studies (Chen *et al.*, 1995; Mariappan *et al.*, 1996) and (Zheng *et al.*, 1996) looked for evidence of quadruplex formation and did not find it. Both studies used Na^+ rather than K^+, reducing their chances of finding stable quadruplexes. Zheng *et al.* (1996) did most of their work on $d(CGG)_3$ and a UV melting-point study with $d(CGG)_4$, so from the findings by Fry & Loeb (1994) and Chen (1995) of very slow formation of quadruplex with $d(CGG)_4$ even in the presence of potassium it is not at all surprising that quadruplexes were not found by Zheng *et al.* (1996). Rather, it is surprising that Kettani *et al.* (1995) did find evidence of quadruplex formation by their short oligonucleotide, and with Na^+. Undoubtedly the three thymidine residues play a part. Chen *et al.* (1995) showed that the equilibrium between duplex and hairpin formation of $d(GGC)_n$ does not go over to mainly hairpin in the presence of 200 mM Na^+ until $n > 7$. Kettani *et al.* (1995) found that $dGCGGT_3GCGG$ did form hairpins. The T_3 would be expected readily to form a loop but not to be helpful to duplex formation with this sequence. It is not obvious, however, why two hairpins of $dGCGGT_3GCGG$ should associate to form a quadruplex though two duplexes of $d(CGG)_3$ or $d(CGG)_4$ do not associate readily to form a quadruplex. The sodium concentrations were similar in the studies by Zheng *et al.* (1996) and Kettani *et al.* (1995), 100 to 150 mM. The DNA concentration used by Kettani *et al.* (1995), ~10 mM of single strands, was higher than that used by Zheng *et al.* (1996), ~0.6 to 2 mM for 1D NMR studies so perhaps this made the difference.

The concentration used by Chen (1995) was very much lower, 40 μM of nucleotides.

The other NMR study (Chen *et al.*, 1995; Mariappan *et al.*, 1996) included $d(GGC)_{3,\ 4,\ 5,\ 6,\ 7\ and\ 11}$, i.e. some longer molecules, but most of these investigations were carried out at a much lower salt concentration (only 10 or 20 mM NaCl, 10 mM phosphate buffer) that may not have been high enough for quadruplex formation. (Their DNA concentrations ranged from 0.6 to 3.3 mM.) Identical imino-proton profiles and temperature dependencies were reported for all of the oligonucleotides under all the conditions used but the only test at a higher Na^+ concentration, a salt and temperature-dependent imino-proton profile conducted between 5 mM and 1 M NaCl, was again performed on a short molecule, $d(GGC)_5$.

We now return to the question of the alignment of folding of the quadruplexes found. Fry & Loeb (1994) assumed bonding in frame 3. Usdin & Woodford (1995) deduced it from their results. Chen (1995) showed that it can occur, but not (at least with $d(CGG)_4$) under physiological conditions. At the time of publication, Usdin & Woodford (1995) were unaware of evidence for C·G·C·G· quartets but they have since mentioned (Weitzmann *et al.*, 1996) the possibility that their quadruplexes may contain a mixture of these and G_4-quartets, i.e. align in frame 1 or 2. We suspect that bonding may be in frame 2. Firstly, the stability of the structures of Usdin & Woodford (1995) at pH 9.3 and 40 mM KCl makes frame 3 seem unlikely. Secondly, whatever the mode of formation of the tetrahelix, it is likely that the first step would be the pairing of two single-stranded parts of the sequence and the evidence is that duplex pairing is in frame 2 (Mitas *et al.*, 1995; Chen *et al.*, 1995; Mariappan *et al.*, 1996; Zheng *et al.*, 1996). Therefore, if the quadruplex is in frame 3, then at some time during formation the alignment has to change from frame 2 to 3. We note that though Usdin & Woodford (1995) deduced that in the absence of potassium $d(CGG)_{20}C$ formed a hairpin in frame 3, their chemical modification gels show that the alignment was in frame 2. The bands corresponding to the 5′ G of each GpG are more intense than those corresponding to the 3′ G residues. This is particularly evident with DMS treatment after BAA treatment and resuspension. At this stage the DNA appears to have been in hairpin form, whatever it had been when initially annealed, and the pattern of guanine bands is just the same as that obtained by Mitas *et al.* (1995).

Could the results reported by Usdin & Woodford (1995) support frame 2 bonding for the quadruplex? Alternative structures for $d(CGG)_{20}C$ in frames 3 and 2 are shown in Figure 5. In the presence of potassium, Usdin & Woodford (1995) found almost complete protection of guanine residues from modification by DMS. If Kettani *et al.* (1995) are right that there are bifurcated hydrogen bonds in the C·G·C·G· quartets, then guanine residues involved in them should be only 50% protected from N7-methylation. However, if as we suspect, the pairing found by O'Brien (1967) and Williams *et al.* (1989) (bonds 2 in Figure 4(a)) applies, all the guanine residues in the stem of the quadruplex would be protected, just as they would in frame 3. This still leaves us to explain why bases in the loops are protected, but frame 3 alignment does not appear to explain this either. Usdin & Woodford (1995) found only C_{11} (the 11th cytosine residue) and no guanine residue to be unprotected from chemical modification. They deduced that C_{11} would be in loop 2 (Figure 5(a)) and surmised that the cytosine residues in loops 1 and 3 might be C·C bonded. This still leaves unexplained why the loop guanine residues were not modified. Usdin and colleagues (Weitzmann *et al.*, 1996) have carried out a detailed investigation of the structural requirements of quadruplexes as detected by their polymerase arrest assay and showed that loops may require at least three bases. As we would not expect the four guanine residues in loops 1 and 3 to be able to form a quartet, we should expect that these bases would at best be only 50% protected from modification (two G·G pairs), and we would not expect the guanine residues in loop 2 to be fully protected either, yet apparently they are all very well protected. In frame 2 (Figure 5(b)) the protection of the cytosine residues in loops 1 and 3 would be explained by C·G bonding but the guanine residues involved in these pairs should be completely unprotected and we might expect a further four to be only 50% protected. We also might expect three guanine residues in loop 2 to be unprotected or only partially protected. However, Usdin and colleagues (Woodford *et al.*, 1994; Howell *et al.*, 1996) have shown that the β^A-globin promoter of *Gallus domesticus* apparently forms a quadruplex structure in which all the loop guanine residues are protected from modification. The authors also suggest that a guanine residue in loop 2 of this quadruplex might be bonded to one in the flanking sequence and we have adopted this suggestion for our frame 2 bonding scheme for $d(CGG)_{20}C$ (Figure 5(b)). Thus, though we do not know why the loop bases should be so well protected, it seems that frame 2 bonding could fit just as well with the data as could frame 3 bonding.

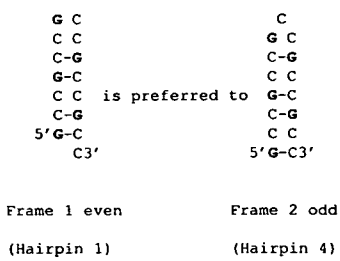
An unresolved question for the frame 2 hypothesis is of the ion-binding. Kettani *et al.* (1995) remarked that it was unclear whether monovalent cation sites could be generated between adjacent quartets in their quadruplex (bonded in frame 2), since the internal coordination sites now consisted of a mixture of favourable guanine O6 oxygen atoms and unfavourable cytosine N4 amino groups. Another way of putting this might be that internally-binding cations might not be required in this case. We suspect, however, that a cation might be coordinated between a G_4-quartet and a C·G·C·G· quartet as there would still be six oxygen atoms at these sites. This would mean that two out of every three inter-quartet sites might be occu-

pied by a cation. We do not envisage that a cation would be coordinated between two C·G·C·G· quartets. The tetraplex of Leonard *et al.* (1995) had no G_4-quartets. It was not found to have a cation between its C·G·C·G· quartets. Having type 2 quartets, it had a narrower narrow groove than the tetraplex described by Kettani *et al.* (1995) and was modelled as being stabilized by a magnesium ion between phosphate groups. Kettani *et al.* (1995) did not report investigation of the ion-dependence or otherwise of their quadruplex. We know only that it can exist, at a very high DNA concentration, in the presence of Na^+ and therefore, presumably, larger quadruplexes in frame 2 could exist in such conditions, and unimolecular quadruplexes might exist at lower DNA concentrations. Our observations on the work of Mitas *et al.* (1995) and Smith *et al.* (1994) discussed earlier, as well as the results of Fry & Loeb (1994) argue that $d(CGG)_n$ quadruplexes can exist in Na^+ solution. Therefore it seems possible that they might be bonded in frame 2 and that their stability might be greatly increased by K^+ binding as we have suggested. It is possible that the structure might be further stabilized by Mg^{2+} in addition to the K^+.

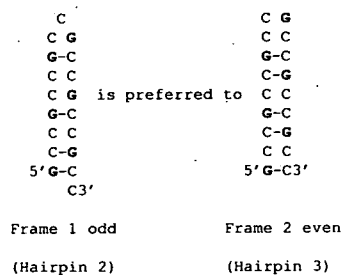
The C-rich Strand: Frame 1 or 2?

Smith *et al.* (1994) found that in native gels both $d(GGC)_{15}$ and $d(CCG)_{15}$ oligonucleotides migrate faster than non-self-complementary markers, suggesting that they adopt unimolecular secondary structures. They showed that the folded C-rich oligonucleotide was a particularly good substrate for human DNA(cytosine-5)methyltransferase. They proposed hairpin formation in frame 1. In this frame the C of CpG is C·C mispaired, an arrangement in which they imply that the C would be more easily methylated than if it were in a C·G bond. They then constrained sequences of $d(CCG)_{11}$ and $d(CGG)_{11}$ to fold in frame 1 by embedding them in flanking sequences that annealed to a complementary oligonucleotide that lacked the sequence to be looped out, and again found that the C-rich sequence was a good substrate for the methyltransferase. We noted, though, that this constrained loop was apparently not as good a substrate as the unconstrained $d(CCG)_{15}$. Further work (Chen *et al.*, 1995; Laayoun & Smith, 1995) confirmed increased methylation of hairpins of the C-rich strand by the same enzyme and this was explained as being due to the increased flexibility of the C·C bond allowing a cytosine residue to be flipped out of the helix more easily.

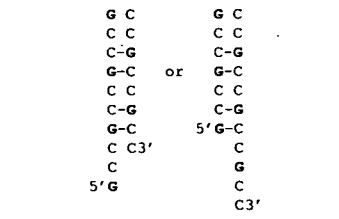
Further electrophoresis of single oligonucleotides showed that the C-rich strand forms hairpins much more easily than does the G-rich strand (Chen *et al.*, 1995). In 5 mM NaCl, hairpin was the predominant form for both strands with 5 to 11 repeats but in 200 mM NaCl $d(GGC)_n$ requires $n > 7$ before hairpin is the dominant form over homoduplex $d(GGC)_n \cdot d(GGC)_n$ and there is still an appreciable proportion in the duplex state at $n = 11$, whereas with $d(GCC)_n$ the hairpin is the dominant form even at $n = 5$. The hairpins were then investigated by NMR (Chen *et al.*, 1995; Mariappan *et al.*, 1996). The results showed that in the C-rich strand the C of the CpG is C·C paired, which indicates frame 1. It was found that $d(GCC)_5$ prefers to pair in frame 1 with an even number of unpaired bases in the loop (i.e. hairpin 1) and an overhanging 3′C rather than pairing in frame 2 with an odd number of unpaired bases in the loop (i.e. hairpin 4) and no overhang, i.e.:



$d(GCC)_6$ was also found to pair in frame 1 with a 3′C overhang. In this case the loop has three unpaired bases (hairpin 2 of our scheme), i.e. it effectively pairs as $dG(CCG)_5CC$. Since a hairpin with this sequence paired with no overhanging base would be a type 3 hairpin, we can infer that hairpin 2 is preferred over hairpin 3, i.e.:

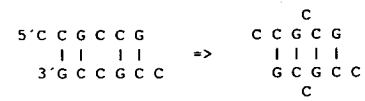


These investigations with short oligonucleotides did not show whether an odd or even-membered loop is preferred in the preferred frame 1. In order for $d(GCC)_6$ to form hairpin 1 it would have to have either a two-base 5′ overhang or a four-base 3′ overhang, i.e.:



and these are presumably energetically less favourable, with such a short stem, than having a hairpin 2 structure and only a one-base 3′ overhang. In a genomic setting, bases not involved in the hairpin would not be overhanging but involved in Watson-Crick pairing with a complementary strand so the preferred loop, whichever it is, would have a better chance of forming.

Another study (Gao *et al.*, 1995) examined $d(CCG)_2$ and several other short C-rich oligonucleotides by NMR to determine the alignment of the homoduplex. The authors concluded that the alignment of $d(CCG)_2 \cdot d(CCG)_2$ is frame 2 with a 5′ overhanging C on each strand and that the single pair of mismatched cytosines residues do not form a C·C bond but that both protrude outwards (and fold in a 5′ direction) on the same side into the minor groove with the C·G pairs on either side stacking upon one another. They named this new kind of duplex DNA the e-motif.



However, investigation of duplexes of d(CGCCG), $d(CGC)_2$ and $d(GCC)_2$ showed no evidence of this structure. Spectra of $d(CCG)_{3\text{-}5}$ were interpreted as showing multiconformational equilibria, which we suspect might include hairpins. The candidates mentioned were the e-motif and parallel duplexes. The spectra of d(CGCCG) and $d(CCG)_{3\text{-}5}$ showed a resonance characteristic of protonated cytosine, thought possibly to indicate $C^+ \cdot C$ bonds. However, the same authors (Zheng *et al.*, 1996) later concluded that in $d(CCG)_{n>2}$ $C^+ \cdot C$ bonds do exist but that the e-motif is in equilibrium with a protonated stacked-in form associated with a regular backbone conformation, and that perterbations in backbone conformation associated with anomalous helical structure appear to be dampened by the dynamic motions of the mismatched bases. The study discussed earlier (Mariappan *et al.*, 1996) did not observe an imino $C^+ \cdot C$ signal. It concluded that the mispaired cytosine residues stacked within the helix but with a single hydrogen bond (which is possible without protonation) that allows them to flip out of the helix more easily than cytosine residues in C·G bonds. There thus appears to be a fair amount of agreement about the mobility of the mispaired cytosine residues; the disagreement is mainly about which cytosine residues they are. Zheng *et al.* (1996) suggested that the e-motif would favour hypermethylation by 5-methyl transferases because the enzymes require cytosine in an extrahelical position. Unfortunately, the cytosine residues in the proposed e-motif, those in the sequence GpC, are not those that are methylated whereas, as has been pointed out (Smith *et al.*, 1994; Mariappan *et al.*, 1996), the cytosine residues mispaired in frame 1, i.e. those of CpG, are methylated. Recent results (Yu *et al.*, 1997), however, have brought a surprise.

Following work on the G-rich strand (Mitas *et al.*, 1995), Yu *et al.* (1997) have investigated secondary structure of the C-rich strand by chemical and enzymic cleavage as well as by physical studies. As before, they considered two alignments, frames 1 and 2, and in these considered quadruplexes and hairpins, and in addition, in both frames, hairpins with all the cytosine residues turned outwards into the minor groove, both referred to as extended e-motif. First they point out that as cytosine has the highest pK_a among all the bases, $d(CCG)_n$ might exhibit pH-dependent structural transitions. An oligonucleotide containing $d(CCG)_{15}$ was studied by electrophoretic mobility, UV absorbance and circular dichroism spectroscopy over ranges of temperature and/or pH and was compared with oligonucleotides containing $d(C^mCG)_{15}$ and other lengths and sequences of repeats in the first study. The conclusion of this is that though there is some protonation of cytosine residues, even at physiological pH, there are no $C^+ \cdot C$ bonds at or above pH 6·5 (and that quadruplexes do not form, unless possibly below this pH where there is evidence of $C^+ \cdot C$ bonds). This directs attention to the extended e-motif possibilities.

Guanine residues were modified with DMS and unpaired cytosine residues with hydroxylamine or 2-hydroperoxytetrahydrofuran and cleaved with piperidine. If the alignment was in frame 1 then the cytosine residues 5′ to the guanine residues would be mispaired but if it was frame 2, then the 3′ cytosine residues would be mispaired. All the guanine residues cleaved, confirming absence of quadruplex formation. As with their G-rich strand investigations (Mitas *et al.*, 1995), the DNA was annealed for five minutes on ice but in this case, as the other evidence suggested that quadruplexes would be unlikely to form, this was not so important. The very interesting and surprise finding was that the vast majority of the cytosine cleavage was of the residues 3′ to the guanine residues, revealing alignment in frame 2. Cleavage 5′ to unpaired bases by P_1 nuclease confirmed this. As in the G-rich strand investigations, the repeat sequence was set in flanking DNA. In case the alignment had been forced by pairing in the flanking DNA or by the stability of the loop, changes were made in the flanking sequence and the number of repeats was increased (to diminish the effect of the loop) and changed to an even number (18 and 20 trinucleotides) to change the loop to that predicted to be more stable in frame 1. Despite all this, the oligonucleotides still annealed in frame 2 and even looped out a base on one side in order to achieve this.

Yu *et al.* (1997) also found that the backbone was distorted. This was indicated by the fact that there was some P_1 cleavage of the backbone between the two adjacent C·G base-pairs. Such cleavage did not occur between the adjacent C·G base-pairs in a hairpin of GTC repeats (whose pairing is analogous to frame 2 of CCG repeats). In both the oligonucleotides containing $d(CCG)_{15}$ and $d(CCG)_{20}$ the nuclease cleaved very strongly on both sides of a single cytosine residue in the hairpin stem, indicating that it was flipped right out of the helix. The conclusion of all the studies was that when $n \geqslant 15$ d(CCG) repeat hairpins pair in frame 2 and adopt the e-motif, and that the unpaired cytosine residues that are turned outwards into the minor groove fold back in a 5′ direction so far as to stack

with another cytosine residue folded towards it in a 5′ direction on the other strand but separated from it by two intervening C·G base-pairs. This model was developed by computer simulation, starting from co-ordinates supplied by X. Gao and colleagues. Further simulation predicted that this stacking causes such stress on the helix as to cause an occasional cytosine residue to be flipped right out. This is in contrast to the conclusion reached by Zheng *et al.* (1996) from their NMR data that the e-motif occurred in duplexes of $d(CCG)_2$ and that, as chain length increased, cytosine residues were in equilibrium between the e-motif and the stacked-in position and backbone distortion was smoothed out, though labile.

Yu *et al.* (1997) addressed two questions arising from their results. First, they accepted the NMR evidence that short $d(GCC)_n$ oligonucleotides ($n = 5$ to 7) form hairpins in frame 1 (Chen *et al.*, 1995; Mariappan *et al.*, 1996). Their explanation is that with short tracts loop structure and/or end effects might favour frame 1. Actually, as we have discussed above, Mariappan *et al.* (1996) showed that frame 1 was preferred for $d(GCC)_n$ oligonucleotides regardless of whether the loop was odd or even-membered and even though frame 1 gave 3′ overhangs in this frame while frame 2 would have given no overhang. The real explanation may be connected with the fact that both NMR investigations (Chen *et al.*, 1995; Mariappan *et al.*, 1996) and (Gao *et al.*, 1995; Zheng *et al.*, 1996) found that there was some C·C bonding with short oligonucleotides, whereas Yu *et al.* (1997) found none with their longer molecules. When the cytosine residues are in the stacked-in position, frame 1 is the more stable because, as pointed out by Yu *et al.* (1997), the stacking energy of the GpC base-pair steps present in the frame 1 alignment is −14.59 kcal/mol as against −9.69 kcal/mol for the CpG steps present in frame 2. However, with the cytosine residues turned outwards frame 2 is more stable because now the stacking of the base-pairs on either side of the outwardly turned cytosine residues is more critical and, as pointed out by Yu *et al.* (1997), this is a pseudo-GpC step in frame 2 but a pseudo-CpG step in frame 1. The extended e-motif is, one might say, "locked" by the stacking of cytosine residues 3 bp apart reaching towards each other in the minor groove. In short oligonucleotides, perhaps, there are not enough of these stacked cytosine pairs to stabilize the structure. For instance, in a hairpin of $d(GCC)_7$ there can be only two such pairs. We have suggested (Darlow & Leach, 1998) that outwardly turned cytosine residues may provide a mechanism by which short hairpins aligned in frame 1, forming in a long tract of repeats, might convert to a frame 2 alignment as more repeats become involved.

The other question was why $d(CCG)_{15}$ hairpins should be a good substrate for 5-methylation of cytosine residues when in frame 2 the wrong cytosine residues are extrahelical. They proposed two possible solutions. Either there might be a minor population of hairpins in frame 1 or the CpG dinucleotides in the extended e-motif frame 2 hairpin might be an excellent substrate for the human methylase because of the distortion of the backbone, as suggested by results reported by Laayoun & Smith (1995). The latter might be the explanation to the observation that the data of Smith *et al.* (1994) showed that $d(CCG)_{11}$ constrained to pair in frame 1 was not quite as good a substrate for methylation as the unconstrained $d(CCG)_{15}$ (after taking account of the difference in number repeats and of the extra non-repeat DNA required for constraint of the hairpin). However, we note that the gels of Yu *et al.* (1997) did show some cleavage of the cytosine residues 5′ of the guanine residues, consistent with a minor population of hairpins in frame 1.

A question not addressed was why frame 1 was determined to be the alignment of short oligonucleotides by Chen *et al.* (1995) and Mariappan *et al.* (1996) but frame 2 by Gao *et al.* (1995) and Zheng *et al.* (1996). The former used oligonucleotides of five to seven trinucleotides that were long enough to form hairpins. They were $d(GCC)_n$, which might have been expected to align in frame 2, yet did not. ($d(CCG)_{5\text{-}7}$ and $d(CGC)_{5\text{-}7}$ were not examined to see whether they would also adopt frame 1.) The major investigations of Gao *et al.* (1995) were on $d(CCG)_2$ oligonucleotides that were too short to form hairpins and so formed duplexes. They might have been expected to align in frame 1, but the data were interpreted as showing the e-motif in frame 2. The investigations of $d(GCC)_2$ and $d(CGC)_2$ were only of the 1D exchangeable proton resonance spectra. They did not support an e-motif interpretation but the alignment was not investigated further.

Conclusions

We hope that this detailed analysis of the apparently conflicting papers on secondary structure formation by the single strands of d(CGG)·d(CCG) repeat tracts has cleared up some of the confusion. The G-rich strand forms hairpins in frame 2 but will also undoubtedly form quadruplexes. $d(CGG)_4$ can form parallel-stranded tetramolecular quadruplexes. These have the same base-pairing arrangement as would an antiparallel quadruplex in frame 3 but formation is very slow at acid pH and extremely slow at pH 8. However, tracts of 13 or more trinucleotides can form unimolecular (antiparallel) quadruplexes very rapidly at 37°C and these may be bonded in frame 2. These quadruplexes, however, were observed at pH values of 9 or more. We have discussed evidence that such quadruplexes can form at pH values as low as 8.3 and consider that d(CGG) repeats may form quadruplexes under physiological conditions. Evidence appears to suggest that the C-rich strand forms hairpins in frame 1 with the mismatched cytosine residues stacked into the helix in hairpins of up to

at least seven trinucleotides but with 15 or more trinucleotides (and possibly less) it forms distorted hairpins in frame 2 with the mispaired cytosine residues turned into the minor groove in the extended e-motif. Our own results (reported elsewhere in this issue) suggest that, *in vivo*, small hairpins can form in frame 1 or 2, and that in frame 1 an even-membered loop is energetically more favourable (as found *in vitro* by Mariappan *et al.*, 1996) while in frame 2 an odd-membered loop is favoured. The demonstration that these structures can form *in vivo* lends support to ideas that they could be involved in the expansion process. Finally, it seems worth quoting the words of Lee (1990), written before the discovery of trinucleotide repeat expansion. "*In vitro*, Na^+, K^+, Mg^{2+} and Ca^{2+} are all present in the cytoplasm and presumably the nucleus of eukaryotic cells. It is known that the Ca^{2+} concentration, for example, increases dramatically upon fertilization of an oocyte. Therefore the structure which is adopted by the guanine-rich telomers", and here we might add d(CGG) repeats, "may change during the cell cycle and will be dependent on a subtle balance between the concentrations of Na^+ and K^+ on the one hand and Mg^{2+} and Ca^{2+} on the other." This dependence of quadruplex formation upon cation concentrations could turn out to be a mechanism behind differences in expansion frequency between stages in the life-cycle, tissues and sexes (Ashley & Warren, 1995). As far as we are aware, no report of the effect of calcium upon the stability of secondary structure in d(CGG) repeats has yet been published.

Acknowledgements

J. M. D. was supported by a Medical Research Council Human Genome Mapping Project grant. D. R. F. L. is supported by grants from the Medical Research Council, the Biotechnology and Biological Sciences Research Council and The Wellcome Trust.

References

Acevedo, O. L., Dickenson, L. A., Macke, T. J. & Thomas, C. A., Jr (1991). The coherence of synthetic telomeres. *Nucl. Acids Res.* **19**, 3409–3419.

Ashley, C. T. & Warren, S. T. (1995). Trinucleotide repeat expansion and human disease. *Annu. Rev. Genet.* **29**, 703–728.

Chen, F.-M. (1995). Acid-facilitated supramolecular assembly of G-quadruplexes in $d(CGG)_4$. *J. Biol. Chem.* **270**, 23090–23096.

Chen, X., Mariappan, S. V. S., Catasti, P., Ratliff, R., Moyzis, R. K., Laayoun, A., Smith, S. S., Bradbury, E. M. & Gupta, G. (1995). Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. *Proc. Natl Acad. Sci. USA*, **92**, 5199–5203.

Darlow, J. M. & Leach, D. R. F. (1998). Evidence for two preferred hairpin folding patterns in d(CGG)·d(CCG) repeat tracts *in vivo*. *J. Mol. Biol.* **275**, 17–23.

Fry, M. & Loeb, L. A. (1994). The fragile X syndrome $d(CGG)_n$ nucleotide repeats form a stable tetrahelical structure. *Proc. Natl Acad. Sci. USA*, **91**, 4950–4954.

Gacy, A. M., Goellner, G., Juranic, N., Macura, S. & McMurray, C. T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. *Cell*, **81**, 533–540.

Gao, X., Huang, X., Kenneth, S. G., Zheng, M. & Liu, H. (1995). New antiparallel duplex motif of DNA CCG repeats that is stabilised by extrahelical bases symmetrically located in the minor groove. *J. Am. Chem. Soc.* **117**, 8883–8884.

Hammond-Kosack, M. C. U., Kilpatrick, M. W. & Docherty, K. (1992). Analysis of DNA structure in the human insulin gene-linked polymorphic region *in vivo*. *J. Mol. Endocrinol.* **9**, 221–225.

Hardin, C. C., Henderson, E., Watson, T. & Prosser, J. K. (1991). Monovalent cation induced structural transitions in telomeric DNAs: G-DNA folding intermediates. *Biochemistry*, **30**, 4460–4472.

Hardin, C. C., Watson, T., Corregan, M. & Bailey, C. (1992). Cation-dependent transition between the quadruplex and Watson-Crick hairpin forms of $d(CGCG_3GCG)$. *Biochemistry*, **31**, 833–841.

Hardin, C. C., Corregan, M., Brown, B. A. I. & Frederick, L. N. (1993). Cytosine-cytosine$^+$ base pairing stabilizes DNA quadruplexes and cytosine methylation greatly enhances the effect. *Biochemistry*, **32**, 5870–5880.

Henderson, E. R., Moore, M. & Malcolm, B. A. (1990). Telomere G-strand structure and function analysed by chemical protection, base analogue substitution, and utilization by telomerase *in vitro*. *Biochemistry*, **29**, 732–737.

Howell, R. M., Woodford, K. J., Weitzmann, M. N. & Usdin, K. (1996). The chicken β-globin gene promoter forms a novel "cinched" tetrahelical structure. *J. Biol. Chem.* **271**, 5208–5214.

Kettani, A., Kumar, R. A. & Patel, D. J. (1995). Solution structure of a quadruplex syndrome triplet repeat. *J. Mol. Biol.* **254**, 638–656.

Kettani, A., Bouaziz, S., Wang, W., Jones, R. A. & Patel, D. J. (1997). *Bombyx mori* single repeat telomeric DNA sequence forms a G-quadruplex capped by base triads. *Nature Struct. Biol.* **4**, 382–389.

Kubitschek, H. E. & Henderson, T. R. (1966). DNA replication. *Proc. Natl Acad. Sci. USA*, **55**, 512–519.

Kuryavyi, V. V. & Jovin, T. M. (1995a). Triad-DNA. *J. Biomol. Struct. Dynam.* **12**, a126.

Kuryavyi, V. V. & Jovin, T. M. (1995b). Triad-DNA: a model for trinucleotide repeats. *Nature Genet.* **9**, 339–341.

Laayoun, A. & Smith, S. S. (1995). Methylation of slipped duplexes, snapbacks and cruciforms by human DNA(cytosine-5)methyltransferase. *Nucl. Acids Res.* **23**, 1584–1589.

Lee, J. S. (1990). The stability of polypurine tetraplexes in the presence of mono- and divalent cations. *Nucl. Acids Res.* **18**, 6057–6060.

Leonard, G. A., Zhang, S., Peterson, M. R., Harrop, S. J., Helliwell, J. R., Cruse, W. B. T., Langlois, d'. Estaintot B., Kennard, O., Brown, T. & Hunter, W. N. (1995). Self-association of a DNA loop creates a quadruplex: crystal structure of d(GCATGCT) at 1.8 Å resolution. *Structure*, **3**, 335–340.

Löwdin, P.-O. (1964). Some aspects on DNA replication; incorporation errors and proton transfer. In *Electronic Aspects of Biochemistry* (Pullman, B., ed.), pp. 167–201, Academic Press, New York and London.

Mariappan, S. V. S., Catasti, P., Chen, X., Ratliff, R., Moysis, R. K., Bradbury, E. M. & Gupta, G. (1996). Solution structures of the individual single strands of the fragile X DNA triplets $(GCC)_n \cdot (GGC)_n$. *Nucl. Acids Res.* **24**, 784–792.

McGavin, S. (1971). Models of specifically paired like (homologous) nucleic acid structures. *J. Mol. Biol.* **55**, 293–298.

Mitas, M., Yu, A., Dill, J. & Haworth, I. S. (1995). The trinucleotide repeat sequence $d(CGG)_{15}$ forms a heat-stable hairpin containing $G^{syn} \cdot G^{anti}$ base pairs. *Biochemistry*, **34**, 12803–12811.

Nadel, Y., Weisman-Shomer, P. & Fry, M. (1995). The fragile X syndrome single strand $d(CGG)_n$ nucleotide repeats readily fold back to form unimolecular hairpin structures. *J. Biol. Chem.* **270**, 28970–28977.

O'Brien, E. J. (1967). Crystal structures of two complexes containing guanine and cytosine derivatives. *Acta Crystallog.* **23**, 92–106.

Sen, D. & Gilbert, W. (1990). A sodium-potassium switch in the formation of four-stranded G4-DNA. *Nature*, **344**, 410–414.

Sinden, R. R. & Wells, R. D. (1992). DNA structure, mutations and human genetic disease. *Curr. Opin. Biotechnol.* **3**, 612–622.

Smith, S., Laayoun, A., Lingeman, R. G., Baker, D. J. & Riley, J. (1994). Hypermethylation of telomere-like foldbacks at codon 12 of the human c-Ha-*ras* gene and the trinucleotide repeat of the *FMR-I* gene of fragile X. *J. Mol. Biol.* **243**, 143–151.

Sutherland, G. R. & Richards, R. I. (1995). The molecular basis of fragile sites in human chromosomes. *Curr. Opin. Genet. Dev.* **5**, 323–327.

Usdin, K. & Woodford, K. J. (1995). CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis *in vitro*. *Nucl. Acids Res.* **23**, 4202–4209.

Venczel, E. A. & Sen, D. (1993). Parallel and antiparallel G-DNA structures from a complex telomeric sequence. *Biochemistry*, **32**, 6220–6228.

Weitzmann, M. N., Woodford, K. J. & Usdin, K. (1996). The development and use of a DNA polymerase arrest assay for the evaluation of parameters affecting intrastrand quadruplex formation. *J. Biol. Chem.* **271**, 20958–20964.

Williams, N. G., Williams, L. D. & Shaw, B. R. (1989). Dimers, trimers, and tetramers of cytosine with guanine. *J. Am. Chem. Soc.* **111**, 7205–7209.

Williamson, J. R. (1994). G-quartet structures in telomeric DNA. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 703–730.

Woodford, K. J., Howell, R. M. & Usdin, K. (1994). A novel K^+-dependent DNA synthesis arrest site in a commonly occuring sequence motif in eukaryotes. *J. Biol. Chem.* **269**, 27029–27035.

Yu, A., Barron, M. D., Romero, R. M., Christy, M., Gold, B., Jianli, D., Gray, D. M., Haworth, I. S. & Mitas, M. (1997). At physiological pH $d(CCG)_{15}$ forms a hairpin containing protonated cytosines and a distorted helix. *Biochemistry*, **36**, 3687–3699.

Zheng, M., Huang, X., Smith, G. K., Yang, X. & Gao, X. (1996). Genetically unstable CXG repeats are structurally dynamic and have a high propensity for folding. An NMR and UV spectroscopic study. *J. Mol. Biol.* **264**, 326–336.

Edited by I. Tinoco

(Received 20 June 1997; received in revised form 23 September 1997; accepted 26 September 1997)

J. Mol. Biol. (1998) 275, 17–23

JMB

COMMUNICATION

Evidence for Two Preferred Hairpin Folding Patterns in d(CGG)·d(CCG) Repeat Tracts *in vivo*

John M. Darlow and David R. F. Leach*

Institute of Cell and Molecular Biology, University of Edinburgh, King's Buildings Edinburgh EH9 3JR, UK

*Corresponding author

Unusual DNA secondary structures have been implicated in the expansion of trinucleotide repeat tracts that has been found to be responsible for a growing number of human inherited disorders and folate-sensitive fragile chromosome sites. By inserting trinucleotide repeat sequences into a palindromic clamp in λ phage we are able to investigate their tendencies to form hairpins *in vivo* in any particular alignment and with odd or even numbers of repeat units in the hairpin. We previously showed that with d(CAG)·d(CTG) repeat tracts there was a markedly greater tendency to form hairpins with even numbers of repeat units than with odd numbers, whereas d(GAC)·d(GTC) repeats showed no such alternation despite having the same base composition. We expected that d(CGG)·d(CCG) repeats, might show the same pattern as d(CAG)·(CTG) repeats since they are also involved in trinucleotide repeat expansion disorders The pattern was not so clear and we wondered whether this might be because d(CGG)·d(CCG) repeats have more than one possible alignment in which they could self-anneal. We now present results for all three alignments, which suggest that while even-membered hairpins are preferred in the frame d(CGG)·d(CCG), hairpins with odd numbers of trinucleotides are more stable in the frame d(GGC)·d(GCC). In both cases the base-pair predicted to close the terminal loop of unpaired bases is 5′C·3′G which has previously been found to be a favoured loop-closing pair.

Keywords: trinucleotides; CGG repeats; DNA hairpins; palindrome; *in vivo*

DNA sequences of multiple repeats are prone to mutations altering the number of repeat units (Levinson & Gutman, 1987) but some trinucleotide repeat sequences are particularly unstable with very high mutation rates. In the last six years an increasing number of human genetic disorders and fragile chromosome sites has been found to be due to expansion of specific polymorphic trinucleotide repeat tracts (Warren, 1996). So far, the trinucleotide repeated at all but one of these loci has had the sequence CXG, which has led many workers to suggest that single strands of these tracts might be able to form intramolecular "pseudo"-hairpins with C·G and/or G·G bonds, or even quadruplex structures, and that these structures might be instrumental in the expansion mechanism. Evidence for the formation of such structures *in vitro* has been found by a variety of biophysical methods, including electrophoretic migration and nuclear magnetic resonance studies, chemical and enzymatic probing, DNA base-modifications and cleavage, as well as by energy considerations (Kohwi *et al.*, 1993; Mitas *et al.*, 1995; Gacy *et al.*, 1995; Mitchell *et al.*, 1995; Yu *et al.*, 1995a,b; Yu *et al.*, 1995b; Smith *et al.*, 1995; Mariappan *et al.*, 1996a; Petruska *et al.*, 1996; Zheng *et al.*, 1996 for d(CAG)·d(CTG) repeats, and see our review elsewhere in this issue for d(CGG)·d(CCG) repeats). The presence of these or other non-*B* DNA conformations has also been inferred from DNA polymerase pausing *in vitro* and *in vivo* in CGG repeat tracts, and by differences in expansion and contraction of CTG repeat tracts in *Escherichia coli* on the leading and lagging strands of replication and in mismatch-repair-competent and deficient cells (Wells, 1996; Ji *et al.*, 1996).

We have investigated the tendency of trinucleotide repeat tracts to form single-stranded hairpin loops *in vivo* by inserting trinucleotide repeats into the centre of a long palindrome in λ phage (Darlow & Leach, 1995; this work). In wild-type *E. coli*, vectors containing palindromes of over

0022-2836/98/010017-07 $25.00/0/mb971452

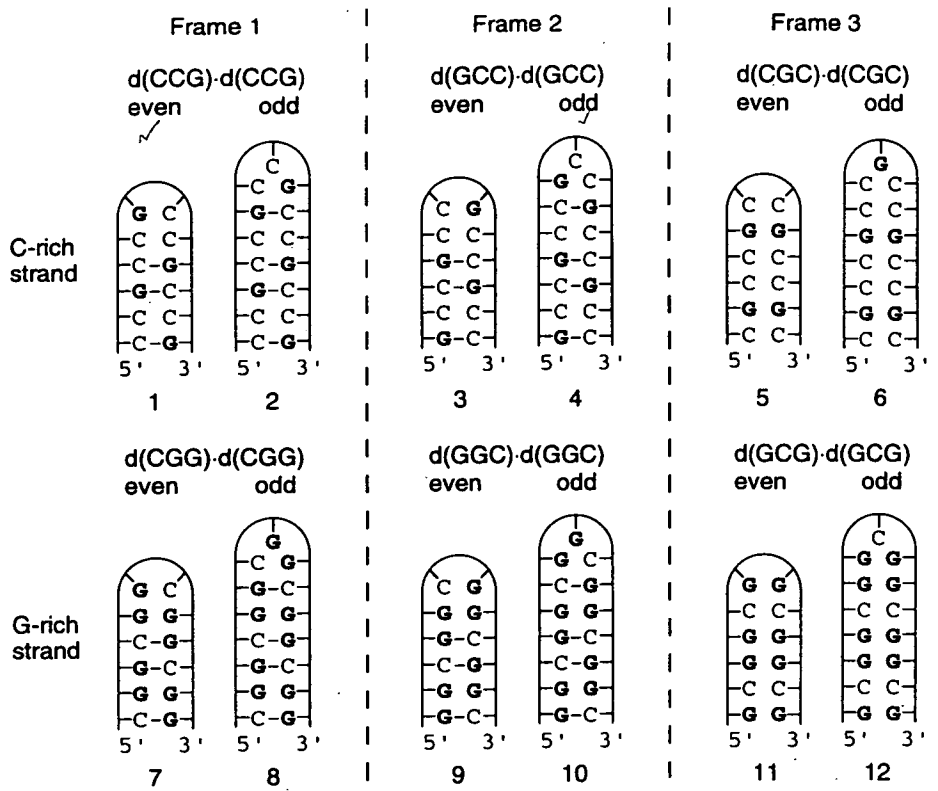


Figure 1. All the possible types of hairpin loops that might be formed by single strands of d(CGG)·d(CCG) repeats. Only Watson-Crick bonds are shown but other workers have found evidence of G·G bonds and C·C bonds *in vitro* (see the text). N.B. The above classification is based upon the alignment of the two sides of the hairpin and the presence of an odd or even number of unpaired bases in the loop. It does not depend upon the actual number of unpaired bases in the loop, length of the stem or the 5'-base of the sequence. We define the alignment by the frame in which the sequence 5'-3' is the same on both sides of the stem. For all three alignments some workers have postulated that long hairpins might fold over to form unistrand quadruplexes.

about 150 to 200 bp are inviable (Leach, 1994). Our assay of trinucleotide repeats is based on the finding that, in an *sbcC* mutant *E. coli* host that will permit long palindrome replication, plaques of phage with palindromes are smaller because of the tendency to form hairpin or cruciform structures that hinder replication. This tendency may be enhanced or reduced by central inserts depending on their ability to form hairpins, and the plaque size is correspondingly reduced or increased (Davison & Leach, 1994a,b). Because hairpins or cruciforms will form only if the palindrome folds in its centre, we are able to determine the position of folding within a repeat tract by varying the number and frame of the repeats inserted. Thus, if the insert is in the frame $d(CGG)_n \cdot d(CCG)_n$ then the palindrome can form a hairpin or cruciform only if individual strands align as $d(CGG)_n \cdot d(CGG)_n$ or $d(CCG)_n \cdot d(CCG)_n$.

We have previously shown (Darlow & Leach, 1995) that d(CAG)·d(CTG) repeats give an alternating pattern, with even numbers of repeat units giving smaller plaques than odd numbers, with $d(CAG)_2 \cdot d(CTG)_2$ apparently forming a particularly tight loop. In contrast, d(GAC)·d(GTC) repeats, which are not associated with repeat expansion disorders, showed a steadily increasing plaque size with increasing numbers of repeats. We had expected that d(CGG)·d(CCG) repeats might show the same pattern as d(CAG)·d(CTG) repeats but found that beyond three units the pattern was different. It occurred to us that this might be because $d(CGG)_n \cdot d(CCG)_n$ tracts have potential for folding in more than one alignment, while d(CAG)·d(CTG) tracts have only the one alignment in which folding is likely.

Figure 1 shows all the possible types of hairpins that might be formed by either strand of $d(CGG)_n \cdot d(CCG)_n$ tracts in all three possible alignments. Frame 1 is the one in which we had previously tested the sequence. Frame 2 is analogous to a $d(GAC)_n \cdot d(GTC)_n$ tract. In frame 3 no Watson-Crick bonding is possible but hairpin-like pairing in this frame with G·G Hoogsteen bonds had been suggested (Sinden & Wells, 1992) and evidence has been put forward for its occurrence *in vitro* (Usdin & Woodford, 1995; Nadel *et al.*, 1995). We now present results of plaque assays of all three alignments, including a repeat of the frame 1 assay under the same conditions used for the other two alignments, and discuss these results in relation to our scheme of Figure 1 and *in vitro* evidence for the different types of hairpins.

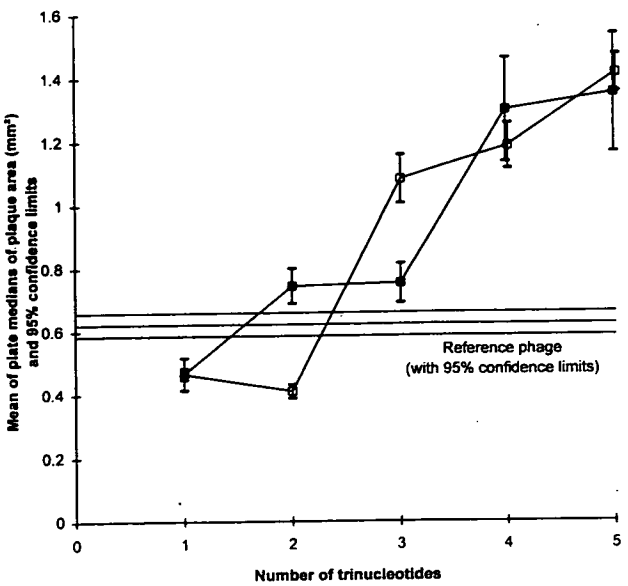


Figure 2. Plaque areas of bacteriophage plotted against number of inserted trinucleotides of d(CGG)·d(CCG) (□) and d(GGC)·d(GCC) (■). The overlapping error bars at four and five repeats have been offset.

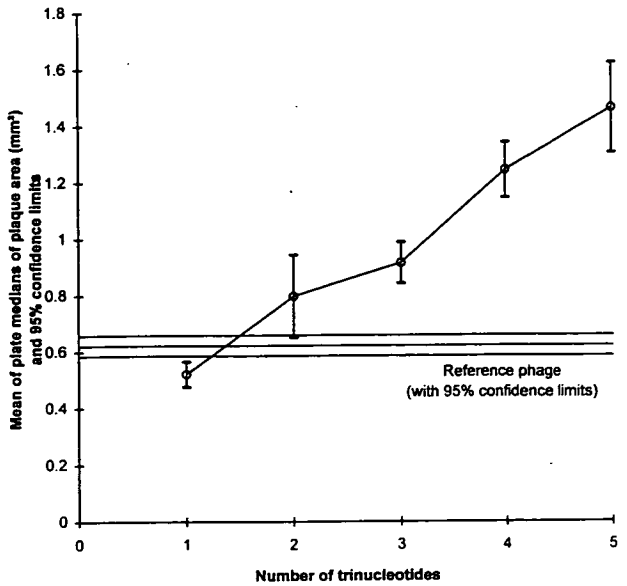


Figure 3. Plaque areas of bacteriophage plotted against number of inserted d(GCG)·d(CGC) trinucleotides.

The results for the $d(CGG)_n \cdot d(CCG)_n$ and $d(GGC)_n \cdot d(GCC)_n$ phage are plotted together in Figure 2 and for the $d(GCG)_n \cdot d(CGC)_n$ phage in Figure 3. In all of these cases, as for $d(CAG)_n \cdot d(CTG)_n$ and $d(GAC)_n \cdot d(GTC)_n$ (Darlow & Leach, 1995), increasing length of non-palindromic DNA between the two inverted repeats that constitute the long palindrome lead to an increase in plaque size. In the absence of any tendency to hairpin formation in the inserted sequence one might expect the graph of plaque area against length of insert to be a straight line or a smooth curve reaching a plateau when the inverted repeat sequences had been separated so far as to have a negligible effect upon phage replication. In fact, inserts in frames 1 and 2 of odd and even numbers of the trinucleotide repeat investigated here produce alternating patterns of plaque area in opposite directions (Figure 2). Smaller plaques than expected suggest a tendency to form a hairpin with its fold in the centre of the inserted sequence and larger plaques than expected suggest a tendency not to form a central fold. Thus, our results suggest that even-membered hairpins are preferred in frame 1 (i.e. type 1 or 7 in Figure 1) and odd-membered hairpins are preferred in frame 2 (i.e. type 4 or 10). In all cases the Watson-Crick base-pair predicted to close the loop of unpaired bases is 5′C·3′G, an arrangement that has previously been shown to be particularly favourable *in vitro* (Hilbers *et al.*, 1994) and *in vivo* (Davison & Leach, 1994b). (The loops in hairpin types 2, 3, 8 and 9 are closed by 5′G·3′C.) It is interesting that we do see a zig-zag pattern in frame 2, $d(GGC)_n \cdot d(GCC)_n$, though we found no such pattern in d(GAC)·d(GTC) repeats (Darlow & Leach, 1995).

Our assay depends upon the finding that the plaque size in palindrome-bearing phage is acutely sensitive to changes in the central sequence of the palindrome. This suggested that a process similar to "S-type cruciform extrusion" occurs *in vivo* (Davison & Leach, 1994a). In S-type cruciform extrusion (first described by Lilley (1985), who then referred to it as Pathway B and Mechanism B) DNA melting at the centre of the palindrome is followed by formation of a small "protocruciform" and then branch migration results in the involvement of the whole of the palindromic sequence to make a larger cruciform structure. The finding that in all positions outside the central two base-pairs of a palindrome C and G produced smaller plaques than A and T suggested that it is the stability of the protocruciform rather than the tendency to central melting that is the more important in cruciform extrusion *in vivo* (Davison & Leach, 1994a). In determination of the folding potential of trinucleotide repeat sequences, therefore, our assay is most useful to identify the folding position(s) that lead to the formation of the most stable hairpin(s) formed from a small number of trinucleotides. We envisage that a small pseudo-hairpin could be a nucleating structure that could extend to form more complicated secondary structures in longer tracts of the repeats, just as a protocruciform can extend to form a much larger cruciform in a palindromic sequence. The general tendency for plaque size to increase with increasing length of trinucleotide repeat tract inserted suggests that the assay may not be suitable for detection of larger structures. There are two potential reasons why long arrays of trinucleotides might not result in the for-

mation of small plaques despite their ability to fold. The first is that the stability of the trinucleotide repeat pseudo-hairpins, may be lower than fully base-paired hairpins and the duplex-hairpin equilibrium may be more favourable to the duplex for trinucleotide repeats than for palindromes. The second is that if a sequence containing multiple repeats can form a pseudo-hairpin then, as the number of repeats increases, the number of identical copies of the sequence forming the most stable loop also increases and most of these loop positions are not in the centre of the surrounding perfect palindrome into which the repeats have been inserted. Any tendency of the insert to fold stably in a position that is not in the centre of the perfect palindrome will not facilitate its extrusion. Thus, even if a repeat sequence can form stable pseudo-hairpins, a structure will be detected only when it is centrally located with respect to the palindromic arms. Our observations therefore do not argue against the formation of large secondary structures in long arrays of trinucleotide repeats.

Figure 3 shows the results for frame 3. Here, the line is almost straight even with small numbers of repeats, which suggests that there is not much tendency to hairpin formation in this alignment. We cannot rule out the possibility that long tracts of the G-rich strand could form stable secondary structures in this alignment, in particular quadruplexes, but evidence for such structures could also be explained by folding in frame 2 (see Darlow & Leach, 1998).

Which are the most stable types of hairpin?

When the central sequence is not a perfect palindrome, the two strands are different and it is always likely that one of them will form a more stable hairpin than the other. Previous work in this laboratory has indicated that it is the strand that forms the tighter loop that primarily determines the plaque size, as was illustrated in correlation of plaque sizes with loops whose numbers of unpaired bases were known from physical studies (Davison & Leach, 1994b). For d(CAG)·d(CTG) repeats, both strands are expected to self-anneal in the same alignment, i.e. d(CAG)·d(CAG) and d(CTG)·d(CTG), and plaque-assays with this repeat (Darlow & Leach, 1995) gave a clear result indicating that hairpins with even numbers of repeats were more stable than hairpins with odd numbers of repeat units. Our result for d(CGG)·d(CCG) repeats (frame 1) was less clear and we suggested (Darlow & Leach, 1995) that this might have been because the two strands had different folding preferences. In our *in vivo* assay we can test only the folding of both strands at once in the same alignment but we can test the potential to fold in all three alignments and with an odd or even number of trinucleotide units in the hairpin. We have now tested folding in all three alignments of d(CGG)·d(CCG) repeats. The results suggest that pairing with an even number of trinucleotide units in the hairpin is favoured in frame 1 but that pairing with an odd number of units in the hairpin is preferred in frame 2.

From this study alone one cannot tell whether it is the C-rich strand or the G-rich strand that makes the more-stable hairpins in these alignments. One might guess that it would be the C-rich strand because the cytosine residues could stack better into the helix or that it could be the G-rich strand if G·G Hoogsteen bonds form. The frame 1 assay shows that either hairpin 1 is more stable than hairpin 2 or that hairpin 7 is more stable than hairpin 8, the frame 2 assay that either hairpin 4 is more stable than hairpin 3 or that hairpin 10 is more stable than hairpin 9.

In vitro studies of secondary structures formed by each of the complementary strands have been conflicting but the evidence is now overwhelming that hairpins of the G-rich strand adopt frame 2 (Darlow & Leach, 1998 and references therein) and an *in vitro* study (Mariappan *et al.*, 1996b) has shown that hairpin 10 is favoured over hairpin 9 and that hairpin 1 is favoured over hairpin 4. However, Chen *et al.* (1995) showed by electrophoresis of suspensions of single oligonucleotides that, when annealed in 200 mM NaCl, $d(GGC)_n$ requires $n > 7$ before hairpin is the dominant form over homoduplex $d(GGC)_n \cdot d(GGC)_n$ and there is still an appreciable proportion in the duplex state at $n = 11$, whereas with $d(GCC)_n$ the hairpin is overwhelmingly the dominant form even at $n = 5$. The *in vitro* studies of the C-rich strand suggest that it can adopt alignment in either frame 1 or frame 2; short oligonucleotides, up to at least seven trinucleotides, prefer to form hairpins in frame 1 but by 15 trinucleotides frame 2 is the predominant choice (discussed by Darlow and Leach, 1998). These *in vitro* results argue strongly that the result of our frame 1 assay is due to the tendency to hairpin formation by the C-rich strand and that hairpin 1 is preferred over hairpin 2 *in vivo* as *in vitro*. In hairpin 1 the base-pair that we expected might close the loop of unpaired bases is 5′C·3′G. This has been found to be favoured as a loop-closing pair over 5′G·3′C (which is in the equivalent position in hairpin 2: Hilbers *et al.*, 1994; Davison & Leach, 1994b). In frame 2 it seems that neither strand immediately forms hairpins with small numbers of trinucleotides but probably again it is the C-rich strand that has the greater tendency to hairpin formation, hairpin 4 being favoured over hairpin 3. The loops in both hairpin 4 and hairpin 10 are expected to be closed by 5′C·3′G as opposed to 5′G·3′C in hairpins 3 and 9.

What happens in longer tracts?

Our assay is attuned to testing the folding potential of short DNA sequences *in vivo* and that is what we have concentrated on. We have envisaged that hairpins formed by small numbers of trinucleotides may, if stable enough, extend to form larger secondary structures in long tracts of repeats. If

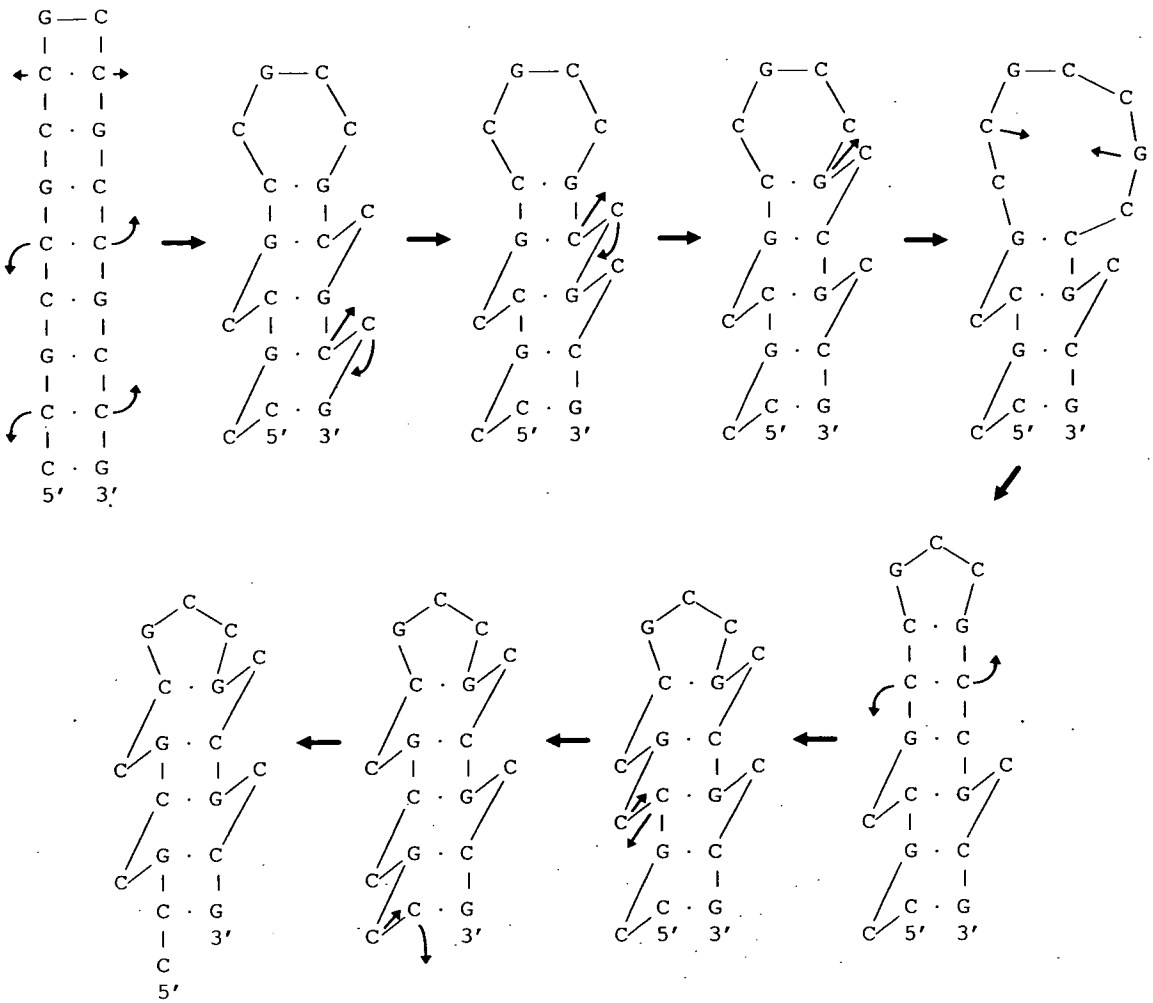


Figure 4. A possible mechanism whereby hairpins of d(CCG) repeats might change their frame of alignment when they reach a length at which frame 2 with the extended e-motif becomes the more stable form. C·C bonds are not very stable, even at low pH, as witnessed by the tendency of the cytosine residues in quadruplexes of $(GCG)_4$ at pH 5.4 to swing outwards with G-tetrads on either side stacking upon one another (Chen *et al.*, 1995). NMR studies of short hairpins and duplexes of $(CCG)_n$ (Chen *et al.*, 1995; Mariappan *et al.*, 1996b; Gao *et al.*, 1995; Zheng *et al.*, 1996) have shown an easy tendency for mismatched cytosine residues to flip out. When the cytosine residues are turned outwards, the adjacent C·G base-pairs may stack. Yu *et al.* (1997) have pointed out that when the cytosine residues are stacked into the helix, frame 1 is the more stable because the stacking energy of the GpC base-pair steps present in that alignment is −14.59 kcal/mol as against −9.69 kcal/mol for the CpG steps present in frame 2. However, when the cytosine residues are turned outwards, frame 2 is more stable because now the stacking of the base-pairs on either side of the outwardly-turned cytosine residues is more critical and this is a pseudo-GpC step in frame 2 but a pseudo-CpG step in frame 1. The top left diagram is of a hairpin in frame 1 and the following diagrams suggest a possible means of transformation to a hairpin in frame 2 with flipped-out cytosine residues (the extended e-motif) at bottom left. The transformation is initiated by the formation of a frame 1 e-motif-like structure that can be converted to a frame 2 e-motif structure by a domino-like exchange of unpaired cytosine residues starting at the 3′ end of the hairpin. (The outwardly-turned cytosine residues fold back in a 5′ direction (Yu *et* al., 1997).)

short hairpins of $d(CCG)_n$ tend to be aligned in frame 1 but longer ones prefer frame 2, we have to explain how a short hairpin with the one alignment could extend to form a longer hairpin with the other alignment. The tendency of mismatched cytosine residues to flip out of the helix (the extended e-motif; Yu *et al.*, 1997; Gao *et al.*, 1995; Darlow & Leach, 1998) may provide a possible mechanism for realignment in the equilibrium between the two alignments because it might allow the change to occur one base-pair at a time with a domino effect (Figure 4). This would be much more energetically feasible than complete melting and reannealing.

Zheng *et al.* (1996) have suggested that the high folding propensity and the dynamic properties that they found in CXG repeats *in vitro* should facilitate formation of local structures, not necessarily hairpins, in competition with a linear duplex in genes containing these repeats. Wells (1996) has drawn diagrams of looped-out structures in different places on the two strands of a trinucleotide repeat

tract that he sees as the only reasonable explanation for the deletion behaviour of such tracts in mismatch-repair deficient bacteria in his laboratory. He called these arrangements slipped structures. Pearson & Sinden (1996) have shown that multiple alternative structures do indeed form in complementary duplex DNA *in vitro* when trinucleotide repeat tracts of disease-causing lengths are melted and reannealed, and have drawn similar diagrams of possible structures, which they call S-DNA. Our results show two different *in vivo* folding preferences of d(CGG)·d(CCG) repeats that may help to generate such structures. How secondary structures might cause dynamic mutation has already been discussed (Darlow & Leach, 1995; Pearson & Sinden, 1996).

Methods

The same two bacteriophage containing long palindromes were used as before (Darlow & Leach, 1995): DRL167 into which inserts containing trinucleotide repeats were ligated, and DRL176 used as a plaque size reference. Two new series of inserts were constructed with central sequences $d(GGC)_n \cdot d(GCC)_n$ and $d(GCG)_n \cdot d(CGC)_n$ by annealing complementary oligonucleotides with the same flanking sequences as before (Darlow & Leach, 1995). Sequences of oligonucleotides were checked by Maxam-Gilbert sequencing and the sizes of the inserts were checked by excising the palindromes by *Eco*RI sites exactly at each end, 3′-end-labelling them and running them on sequencing gels followed by autoradiography.

Plaque size assays were as described (Darlow & Leach, 1995) with the following modifications. Plaque assays were performed on two independent isolates of each of the 15 phage with trinucleotide repeats ($[d(CGG)\cdot d(CCG)]_{1\text{-}5}$, $[d(GGC)\cdot d(GCC)]_{1\text{-}5}$ and $[d(GCG)\cdot d(CGC)]_{1\text{-}5}$) and plaque areas were measured on four plates for each isolate. For the assay of each series of five phage, two stacks of 30 plates were poured. The volume of bottom agar was 42 ml and the plates were left to dry for four days before use. Because drying is always greatest in plates at the top of a stack, the top four plates were set aside and then, taking the first stack, from the fifth plate onwards, plates were dealt in rotation into six piles of four plates, one pile for the first isolate of each of the test phage and one for the reference phage, DRL176, so that each phage had plates from four different parts of the stack. The same was done with the second stack for the second isolate of each phage, again with four plates for the reference phage. If a contaminated plate was encountered, the next plate was taken and dealing went on and the plates used in the other pile were adjusted so that the plates from the same positions were used from each pile.

The plaque areas were measured with an Optimas system (Optimas UK, West Malling, Kent) using Optimas 5.2 software with a Visionplus AFG image capture board and a Pulnix TM-6 monochrome camera run with twin monitors mounted on a Dell XMT 5100 PC clone. At least 60 plaques were measured per plate in two or three fields of view as necessary (up to about 120 plaques). The exact number measured depended upon the density of plaques on the plate as it is felt that to exclude some plaques on the grounds of being surplus could introduce bias. Despite great care, there is always plate-to-plate variation in plaque size, so the plaques of one plate could skew the median just because that plate had by chance more plaques than the others. To avoid this, the median plaque size was determined for each plate and, taking this as the best estimate for the plate, these medians were treated as single datum points and the mean was calculated of all eight results (having first established that both isolates of a phage behaved the same way).

The sizes of plaques of all phage tend to vary between assays and it is not possible to assay all phage at once, hence the use of a reference phage (DRL176). The results from the $[d(GGC)\cdot d(GCC)]_{1\text{-}5}$ and $[d(GCG)\cdot d(CGC)]_{1\text{-}5}$ series were scaled using the reference phage results so that they would be comparable with the $[d(CGG)\cdot d(CCG)]_{1\text{-}5}$ results. The two sets of data were multiplied respectively by 1.46 and 1.19.

Acknowledgements

We thank Chris Jeffrey and John Findlay for facilitating the image analysis and Ewa Okely for technical assistance. J. M. D. was supported by a studentship from the Medical Research Council (MRC) Human Genome Mapping Project and this work is supported by the MRC.

References

Chen, X., Mariappan, S. V. S., Catasti, P., Ratliff, R., Moyzis, R. K., Laayoun, A., Smith, S. S., Bradbury, E. M. & Gupta, G. (1995). Hairpins are formed by the single DNA strands of the fragile X triplet repeats: structure and biological implications. *Proc. Natl Acad. Sci. USA*, **92**, 5199–5203.

Darlow, J. M. & Leach, D. R. F. (1995). The effects of trinucleotide repeats found in human inherited disorders on palindrome inviability in *Escherichia coli* suggest hairpin folding preferences *in vivo*. *Genetics*, **141**, 825–832.

Darlow, J. M. & Leach, D. R. F. (1998). Secondary structures in d(CGG)·d(CCG) repeat tracts. *J. Mol. Biol.* **275**, 3–16.

Davison, A. & Leach, D. R. F. (1994a). The effects of nucleotide sequence changes on DNA secondary structure formation in *Escherichia coli* are consistent with cruciform extrusion *in vivo*. *Genetics*, **137**, 361–368.

Davison, A. & Leach, D. R. F. (1994b). Two-base DNA hairpin-loop structures *in vivo*. *Nucl. Acids Res.* **22**, 4361–4363.

Gacy, A. M., Goellner, G., Juranic, N., Macura, S. & McMurray, C. T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. *Cell*, **81**, 533–540.

Gao, X., Huang, X., Kenneth, S. G., Zheng, M. & Liu, H. (1995). New antiparallel duplex motif of DNA CCG repeats that is stabilised by extrahelical bases symmetrically located in the minor groove. *J. Am. Chem. Soc.* **117**, 8883–8884.

Hilbers, C. W., Heus, H. A., van Dongen, M. J. P. & Wijmenga, S. S. (1994). The hairpin elements of nucleic acid structure: DNA and RNA folding. *Nucl. Acids Mol. Biol.* **8**, 56–104.

Ji, J., Clegg, N. J., Peterson, K. R., Jackson, A. L., Laird, C. D. & Loeb, L. A. (1996). *In vitro* expansion of GGC:GCC repeats: identification of the preferred strand of expansion. *Nucl. Acids Res.* **24**, 2835–2840.

Kohwi, Y., Wang, H. & Kohwi-Shigematsu, T. (1993). A single trinucleotide 5′AGC3′/5′GCT3′, of the triplet-repeat disease genes confers metal-ion-induced non-*B* DNA structure. *Nucl. Acids Res.* **21**, 5651–5655.

Leach, D. R. F. (1994). Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *BioEssays*, **16**, 893–900.

Levinson, G. & Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* **4**, 203–221.

Lilley, D. M. J. (1985). The kinetic properties of cruciform extrusion are determined by DNA base-sequence. *Nucl. Acids Res.* **13**, 1443–1465.

Mariappan, S. V. S., Garcia, A. E. & Gupta, G. (1996a). Structure and dynamics of the DNA hairpins formed by tandemly repeated CTG triplets associated with myotonic dystrophy. *Nucl. Acids Res.* **24**, 775–783.

Mariappan, S. V. S., Catasti, P., Chen, X., Ratliff, R., Moysis, R. K., Bradbury, E. M. & Gupta, G. (1996b). Solution structures of the individual single strands of the fragile X DNA triplets $(GCC)_n \cdot (GGC)_n$. *Nucl. Acids Res.* **24**, 784–792.

Mitas, M., Yu, A., Dill, J., Kamp, T. J., Chambers, E. J. & Haworth, I. S. (1995). Hairpin properties of single-stranded DNA containing a GC-rich triplet repeat: $(CTG)_{15}$. *Nucl. Acids Res.* **23**, 1050–1059.

Mitchell, J. E., Newbury, S. F. & McClellan, J. A. (1995). Compact structures of $d(CNG)_n$ oligonucleotides in solution and their possible relevance to fragile X and related human genetic diseases. *Nucl. Acids Res.* **23**, 1876–1881.

Nadel, Y., Weisman-Shomer, P. & Fry, M. (1995). The fragile X syndrome single strand $d(CGG)_n$ nucleotide repeats readily fold back to form unimolecular hairpin structures. *J. Biol. Chem.* **270**, 28970–28977.

Pearson, C. E. & Sinden, R. R. (1996). Alternative structures in duplex DNA formed within the trinucleotide repeats of the myotonic dystrophy and fragile X loci. *Biochemistry*, **35**, 5041–5053.

Petruska, J., Arnheim, N. & Goodman, M. F. (1996). Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. *Nucl. Acids Res.* **24**, 1992–1998.

Sinden, R. R. & Wells, R. D. (1992). DNA structure, mutations and human genetic disease. *Curr. Opin. Biotechnol.* **3**, 612–622.

Smith, G. K., Jie, J. & Fox, G. E. (1995). DNA CTG triplet repeats involved in dynamic mutations of neurologically related gene sequences form stable duplexes. *Nucl. Acids Res.* **23**, 4303–4311.

Usdin, K. & Woodford, K. J. (1995). CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis *in vitro*. *Nucl. Acids Res.* **23**, 4202–4209.

Warren, S. T. (1996). The expanding world of trinucleotide repeats. *Science*, **271**, 1374–1375.

Wells, R. D. (1996). Molecular basis of genetic instability of triplet repeats. *J. Biol. Chem.* **271**, 2875–2878.

Yu, A., Dill, J., Wirth, S. S., Huang, G., Lee, V. H., Haworth, I. S. & Mitas, M. (1995a). The trinucleotide repeat sequence $d(GTC)_{15}$ adopts a hairpin conformation. *Nucl. Acids Res.* **23**, 2706–2714.

Yu, A., Dill, J. & Mitas, M. (1995b). The purine-rich trinucleotide repeat sequences $d(CAG)_{15}$ and $d(GAC)_{15}$ form hairpins. *Nucl. Acids Res.* **23**, 4055–4057.

Yu, A., Barron, M. D., Romero, R. M., Christy, M., Gold, B., Jianli, D., Gray, D. M., Haworth, I. S. & Mitas, M. (1997). At physiological pH $d(CCG)_{15}$ forms a hairpin containing protonated cytosines and a distorted helix. *Biochemistry*, **36**, 3687–3699.

Zheng, M., Huang, X., Smith, G. K., Yang, X. & Gao, X. (1996). Genetically unstable CXG repeats are structurally dynamic and have a high propensity for folding. An NMR and UV spectroscopic study. *J. Mol. Biol.* **264**, 326–336.

Edited by I. Tinoco

(Received 20 June 1997; received in revised form 23 September 1997; accepted 26 September 1997)